(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
2 August 2001 (02.08.2001)

PCT

(10) International Publication Number
WO 01/55411 A2

(51) International Patent Classification[7]: C12N 15/55, 5/10, 9/16, C07K 16/40, G01N 33/53, C12Q 1/68, A61K 38/46

(21) International Application Number: PCT/US01/03266

(22) International Filing Date: 31 January 2001 (31.01.2001)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
09/495,826 31 January 2000 (31.01.2000) US

(71) Applicant (for all designated States except US): MILLENNIUM PHARMACEUTICALS, INC. [US/US]; 75 Sidney Street, Cambridge, MA 02139 (US).

(72) Inventors; and

(75) Inventors/Applicants (for US only): GLUCKSMANN, Maria, Alexandra [AR/US]; 33 Summit Road, Lexington, MA 02173 (US). WILLIAMSON, Mark [US/US]; 15 Stonecrest Drive, Saugus, MA 01906 (US). RUDOLPH-OWEN, Laura, A. [US/US]; 186 Arborway, #1, Jamaica Plain, MA 02130 (US). TSAI, Fong-Ying [—/US]; 15 Montclair Road, Newton, MA 02468 (US).

(74) Agents: KRON, Eric, J. et al.; Alston & Bird LLP, P.O. Drawer 34009, Charlotte, NC 28234-4009 (US).

(81) Designated States (national): AE, AG, AL, AM, AT, AT (utility model), AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CR, CU, CZ, CZ (utility model), DE, DE (utility model), DK, DK (utility model), DM, DZ, EE, EE (utility model), ES, FI, FI (utility model), GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SK (utility model), SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.

(84) Designated States (regional): ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published:
— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

WO 01/55411 A2

(54) Title: 22438, 23553, 25278, AND 26212 NOVEL HUMAN SULFATASES

(57) Abstract: The present invention relates to newly identified human sulfatases. In particular, the invention relates to sulfatase polypeptides and polynucleotides, methods of detecting the sulfatase polypeptides and polynucleotides, and methods of diagnosing and treating sulfatase-related disorders. Also provided are vectors, host cells, and recombinant methods for making and using the novel molecules.

22438, 23553, 25278, AND 26212 NOVEL HUMAN SULFATASES

FIELD OF THE INVENTION

The present invention relates to newly identified human sulfatases. In particular, the invention relates to sulfatase polypeptides and polynucleotides, methods of detecting the sulfatase polypeptides and polynucleotides, and methods of diagnosing and treating sulfatase-related disorders. Also provided are vectors, host cells, and recombinant methods for making and using the novel molecules.

BACKGROUND OF THE INVENTION

The biology and functions of the reversible sulfation pathway catalyzed by human sulfotransferases and sulfatases has been reviewed by Coughtrie *et al.* (*Chemico-Biological Interactions* 109: 3-27 (1998)). This review, summarized below, focuses on the sulfation of small molecules carried out by cytosolic sulfotransferases rather than the sulfation of macromolecules and lipids catalyzed by membrane-associated sulfotransferases.

Sulfation functions in the metabolism of xenobiotic compounds, steroid biosynthesis, and modulating the biological activity and inactivation and elimination of potent endogenous chemicals such as thyroid hormones, steroids and catechols. This pathway is reversible, comprising the sulfotransferase enzymes that cause the sulfation and the sulfatases that hydrolyze the sulfate esters formed by the action of the sulfotransferases. Accordingly, the interplay between these families regulates the availability and biological activity of xenobiotic and endogenous chemicals. The sulfatases, including the arylsulfatases (ARS), are located in lysosomes or endoplasmic reticulum.

The presence of sulfated components depends upon the availability of key members of the sulfate pathway, i.e., substrate and activated sulfate donor molecule (co-substrate) and the balance between sulfation and sulfate conjugate hydrolysis that depends upon the activity and localization of the sulfotransferases and the sulfatases.

Essentially, divalent sulfate is converted to adenosine 5′ phosphosulfate (PAPS) by hydrolysis of ATP. This compound is in turn converted to 3′ phosphoadenosine 5′ phosphosulfate by hydrolysis of ATP to ADP. This compound is then converted to adenosine 3′ 5′ biphosphate concurrently with the formation of 4-nitrophenolsulfate from 4-nitrophenol. An ARS would then cleave the monovalent sulfate from the 4-nitrophenolsulfate to produce the original 4-nitrophenol. This forms the basis for the sulfation system in humans. Over- or under-production of any of these key molecules can result in sulfate-related disorders. For example, the brachymorphic mouse has a connective tissue disorder that results from a defect in PAPS formation that causes undersulfated cartilage proteoglycans.

ARS enzymes and their genes have been associated with specific genetic diseases. ARSA is located in the lysosomes and removes sulfate from sulfated glycolipids. A deficiency of ARSA has been associated with metachromatic leukodystrophy and multiple sulfatase deficiency (MSD). ARSB is located in lysosomes and has, as an endogenous substrate, dermatan sulfate and chondrotin sulfate. A deficiency of ARSB is associated with Maroteaux-Lamy syndrome and MSD. ARSC is located in the endoplasmic reticulum and has, as its endogenous substrate, cholesterol sulfate and steroid sulfates. A deficiency of ARSC is associated with X-linked ichthyosis and MSD. ARSD may be associated with MSD. ARSE has been associated with chondrodysplasia punctata and MSD. ARSF may be associated with MSD. ARSC hydrolyses sulfate esters on a wide range of steroids and cholesterol. ARSs also hydrolyse sulfate conjugates of xenobiotics.

MSD results from an inability to perform a co- or post-translational modification of a cysteine residue to serine semialdehyde (2-oxo-3-propionic acid). This residue is conserved in all eukaryotic sulfatases described by Coughtrie *et al.* ARSC may have a very broad specificity, extending to iodothyronine sulfates and a number of sulfate conjugates of xenobiotic phenols.

The kinetic and catalytic properties of ARS enzymes in isolation, important for understanding substrate specificity and the physical and chemical properties of enzymes and substrates that allow substrate preference, have been characterized recently based on recombinant enzyme systems. For the expression of the human sulfotransferases, COS and V79 cells have been used. Coughtrie *et al.* have constructed and characterized V79

cell lines stably expressing ARSA, ARSB, and ARSC. These cell lines exhibited the expected substrate preferences of the three enzymes among the substrates 4-nitrocatechol sulfate, estrone sulfate, and dehydroepiandrosterone sulfate(DHEAS).

The sulfation of small molecules can be broadly divided into the areas of chemical defense, hormone biosynthesis, and bioactivation. It was originally viewed that sulfation protected against the toxic effects of xenobiotics in that sulfate conjugates are more readily excreted in urine or bile and generally exhibit reduced pharmacological/biological activity relative to the parent compound. Many drugs and other xenobiotics are conjugated with sulfate. Many phenolic metabolites of the cytochrome P450 mono-oxygenase system are excreted as sulfate conjugates.

Further, potent endogenous chemicals, such as steroids and catecholamines are found at high levels as circulating sulfate conjugates. For example, greater than 90% of circulating dopamine exists as the sulfated form. Sulfation is also suggested to play a role in the inactivation of potent steroids such as estrogens and androgens. Accordingly, sulfation is important in metabolism and homeostasis of such compounds in humans.

DHEAS is the major circulating steroid in humans and estrone sulfate is the major estrogen. These chemicals act as precursors of estrogens and androgens. Extremely large quantities of such steroids or estrogens may occur during various stages of development, such as pregnancy. Estrone sulfate is a precursor for β-estradiol synthesis. In breast cancer cells it is hydrolysed by steroid sulfatase (ARSC) to estrone which is then converted to β-estradiol by action of another enzyme. Accordingly, ARSC is important for maintaining active estrogen. It is thus an important therapeutic target for the treatment of breast cancer.

Cholesterol sulfate, synthesized in the skin epidermis, may have a role in keratinocyte differentiation. Accordingly, hydrolysis of cholesterol sulfate by steroid sulfatase may be important in skin formation and differentiation. This is the major organ affected in X-linked ichthyosis caused by mutations in ARSC.

Although sulfation may widely serve to detoxify potent compounds, some sulfate conjugates are more biologically active than the corresponding parent compound. Minoxidil and cicletanine are activated upon sulfation. Further, an inhibitor of ARSC was shown to potentiate the memory enhancing effect of DHEAS. This suggests a role for sulfates and sulfation in the central nervous system.

An important example of bioactivation by means of sulfation, however, occurs with dietary and environmental mutagens and carcinogens. For a large number of these, sulfation is the terminal step in the pathway to metabolic activation. Examples of such chemicals include aromatic amines (including heterocyclic amines) and benzylic alchohols of chemicals such as polycyclic aromatic hydrocarbons, safrole, and estragole.

The sulfatase gene family has been reviewed in Parenti *et al.* (*Current Opinion in Genetics and Development* 7:386-391 (1997)), summarized below.

The sulfatase family of enzymes is functionally and structurally similar. Nevertheless, these enzymes catalyze the hydrolysis of sulfate ester bonds from a wide variety of substrates ranging from complex molecules such as glycosaminoglycans and sulfolipids to steroid sulfates (see also Coughtrie *et al.*, above). Several human genetic disorders result from the accumulation of intermediate sulfate compounds that result from a deficiency of single or multiple sulfatase activities. A subset of sulfatase, ARS, is characterized by the ability to hydrolyze sulfate esters of chromogenic or fluorogenic aromatic compounds such as *p*-nitrocatechol sulfate and 4-methylumbelliferyl sulfate. Desulfation is required to degrade glycosaminoglycans, heparan sulfate, chondroitin sulfate and dermatan sulfate and sulfolipids. Steroid sulfatase differs from other members of the family with respect to subcellular localization. It is localized in the microsomes rather than in lysosomes. Further, ARSD, ARSE, and ARSF are also non-lysosomal, being localized in the endoplasmic reticulum or Golgi compartment.

The natural substrate of ARSA is cerebroside sulfate. Associated diseases are MLD and MSD. The natural substrate of ARSB is dermatan sulfate. The disease associated with this enzyme is MPSVI and MSD. The natural substrate of ARSC/STS is sulfated steroids. Diseases associated with this enzyme are XLI and MSD. The natural substrates of ARSD-F are unknown. The natural substrates of iduronate-2-sulfate sulfatase (IDS) are dermatan sulfate and herparan sulfate. Diseases associated with this enzyme are MPSII and MSD. The natural substrate of galactose 6-sulfatase is keratan sulfate and chondroitin 6-sulfate. Diseases associated with this enzyme include MPSIVA and MSD. The natural substrate of glucosamine-6-sulfatase is heparan sulfate and keratan sulfate. A disease associated with this enzyme is MPSIIID and MSD. The natural substrate of glucuronate-2-sulfatase is heparan sulfate. The natural substrate of glucosamine-3-sulfatase is heparan sulfate.

Sulfatases are activated through conversion of a cysteine residue as described above. The conversion is required for catalytic activity and is defective in MSD. It is likely that all sulfatases undergo the same modification. The substitution of this cysteine was shown to destroy the enzymatic activity of N-acetyl galactosamine-4-sulfatase (ARSB). It has been shown that the modified residue and a metal ion are located at the base of a substrate binding pocket.

Nine human sulfatase genes are known and murine rat, goat, or avian orthologs for some of these have been identified. A high degree of similarity occurs particularly in the amino terminal region which contains accordingly a potential consensus sulfatase signature.

Sulfatases, as discussed above, are associated with human disease. Most sulfatase deficiencies cause lysosomal storage disorders. The mucopolysaccharidoses contain various associations of mental retardation, facial dysmorphisms, skeletal deformities, hepatosplenomegaly, and deformities of soft tissues caused by deficiencies of sulfatases acting on glycosaminoglycans. In metachromatic leukodystrophy, a deficiency of ARSA causes the storage of sulfolipids in the central and peripheral nervous systems, leading to neurologic deterioration. X-linked icythyosis is caused by STS deficiency leading to increased cholesterol sulfate levels. MSD, a disorder in which all sulfatase activities are simultaneously defective, was shown to result from a defect in the co- or post-translational processing of sulfatases.

Accordingly, sulfatases are a major target for drug action and development. Therefore, it is valuable to the field of pharmaceutical development to identify and characterize previously unknown sulfatases. The present invention advances the state of the art by providing previously unidentified human sulfatases.

SUMMARY OF THE INVENTION

Novel sulfatase nucleotide sequences, and the deduced sulfatase polypeptides are described herein. Accordingly, the invention provides isolated sulfatase nucleic acid molecules having the sequences shown in SEQ ID NOS:2, 4, 6, and 8 or in the cDNA deposited with ATCC as Patent Deposit Number _____, PTA-1639, PTA-1846, or _____, respectively ("the deposited cDNA"), and variants and fragments thereof.

It is also an object of the invention to provide nucleic acid molecules encoding the sulfatase polypeptides, and variants and fragments thereof. Such nucleic acid molecules are useful as targets and reagents in sulfatase expression assays, are applicable to treatment and diagnosis of sulfatase-related disorders and are useful for producing novel sulfatase polypeptides by recombinant methods.

The invention thus further provides nucleic acid constructs comprising the nucleic acid molecules described herein. In a preferred embodiment, the nucleic acid molecules of the invention are operatively linked to a regulatory sequence. The invention also provides vectors and host cells for expressing the sulfatase nucleic acid molecules and polypeptides, and particularly recombinant vectors and host cells.

In another aspect, it is an object of the invention to provide isolated sulfatase polypeptides and fragments and variants thereof, including a polypeptide having the amino acid sequence shown in SEQ ID NOS:1, 3, 5 or 7 or the amino acid sequences encoded by the deposited cDNAs. The disclosed sulfatase polypeptides are useful as reagents or targets in sulfatase assays and are applicable to treatment and diagnosis of sulfatase-related disorders.

The invention also provides assays for determining the activity of or the presence or absence of the sulfatase polypeptides or nucleic acid molecules in a biological sample, including for disease diagnosis. In addition, the invention provides assays for determining the presence of a mutation in the polypeptides or nucleic acid molecules, including for disease diagnosis.

A further object of the invention is to provide compounds that modulate expression of the sulfatase for treatment and diagnosis of sulfatase-related disorders. Such compounds may be used to treat conditions related to aberrant activity or expression of the sulfatase polypeptides or nucleic acids.

The disclosed invention further relates to methods and compositions for the study, modulation, diagnosis and treatment of sulfatase related disorders. The compositions include sulfatase polypeptides, nucleic acids, vectors, transformed cells and related variants thereof. In particular, the invention relates to the diagnosis and treatment of sulfatase-related disorders including, but not limited to disorders as described in the background above, further herein, or involving a tissue shown in the figures herein.

In yet another aspect, the invention provides antibodies or antigen-binding fragments thereof that selectively bind the sulfatase polypeptides and fragments. Such antibodies and antigen binding fragments have use in the detection of the sulfatase polypeptide, and in the prevention, diagnosis and treatment of sulfatase related disorders.

The sulfatases disclosed herein are designated as follows: 22438, 23553, 25278, and 26212.

DESCRIPTION OF THE DRAWINGS

Figure 1 shows the 22438 sulfatase cDNA sequence (SEQ ID NO:2) and the deduced amino acid sequence (SEQ ID NO:1). The 22438 sulfatase coding sequence is set forth in SEQ ID NO:11.

Figure 2 shows a 22438 sulfatase hydrophobicity plot. Relative hydrophobic residues are shown above the dashed horizontal line, and relative hydrophilic residues are below the dashed horizontal line. The cysteine residues (cys) and N glycosylation site (Ngly) are indicated by short vertical lines just below the hydropathy trace. The numbers corresponding to the amino acid sequence (shown in SEQ ID NO:1) of 22438 sulfatase are indicated. Polypeptides of the invention include fragments which include: all or a part of a hydrophobic sequence (a sequence above the dashed line); or all or part of a hydrophilic fragment (a sequence below the dashed line). Other fragments include a cysteine residue or as N-glycosylation site.

Figure 3 shows an analysis of the 22438 sulfatase amino acid sequence: αβturn and coil regions; hydrophilicity; amphipathic regions; flexible regions; antigenic index; and surface probability plot.

Figure 4 shows an analysis of the 22438 sulfatase open reading frame for amino acids corresponding to specific functional sites. For the N-glycosylation sites, the actual modified residue is the first amino acid. For cAMP- and cGMP-dependent protein kinase phosphorylation sites, the actual modified residue is the last amino acid. For protein kinase C phosphorylation sites, the actual modified residue is the

first amino acid. For casein kinase II phosphorylation sites, the actual modified residue is the first amino acid. For N-myristoylation sites, the actual modified residue is the first amino acid. In addition, an amidation site is found from about amino acids 56-59, an EGF-like domain cysteine pattern signature found from about amino acids 260-271, and a sulfatase signature is found from about amino acids 129-138.

Figure 5 shows the 23553 sulfatase cDNA sequence (SEQ ID NO:4) and the deduced amino acid sequence (SEQ ID NO:3). The 23553 sulfatase coding sequence is set forth in SEQ ID NO:12.

Figure 6 shows a 23553 sulfatase hydrophobicity plot. Relative hydrophobic residues are shown above the dashed horizontal line, and relative hydrophilic residues are below the dashed horizontal line. The cysteine residues (cys) and N glycosylation site (Ngly) are indicated by short vertical lines just below the hydropathy trace. The numbers corresponding to the amino acid sequence (shown in SEQ ID NO:3) of 23553 sulfatase are indicated. Polypeptides of the invention include fragments which include: all or a part of a hydrophobic sequence (a sequence above the dashed line); or all or part of a hydrophilic fragment (a sequence below the dashed line). Other fragments include a cysteine residue or as N-glycosylation site.

Figure 7 shows an analysis of the 23553 sulfatase amino acid sequence: αβturn and coil regions; hydrophilicity; amphipathic regions; flexible regions; antigenic index; and surface probability plot.

Figure 8 shows an analysis of the 23553 sulfatase open reading frame for amino acids corresponding to specific functional sites. For the N-glycosylation sites, the actual modified residue is the first amino acid. For protein kinase C phosphorylation sites, the actual modified residue is the first amino acid. For casein kinase II phosphorylation sites, the actual modified residue is the first amino acid. For the tyrosine kinase phosphorylation site, the actual modified residue is the last amino acid residue. For N-myristoylation sites, the actual modified residue is the first amino acid. In addition, a sulfatase signature is found from about amino acids 85-97.

Figure 9 shows relative expression of the 23553 sulfatase mRNA in normal and cancerous human tissues.

Figure 10 shows the 25278 sulfatase cDNA sequence (SEQ ID NO:6) and the deduced amino acid sequence (SEQ ID NO:5). The 25278 sulfatase coding sequence is set forth in SEQ ID NO:13.

Figure 11 shows a 25278 sulfatase hydrophobicity plot. Relative hydrophobic residues are shown above the dashed horizontal line, and relative hydrophilic residues are below the dashed horizontal line. The cysteine residues (cys) and N glycosylation site (Ngly) are indicated by short vertical lines just below the hydropathy trace. The numbers corresponding to the amino acid sequence (shown in SEQ ID NO:5) of 25278 sulfatase are indicated . Polypeptides of the invention include fragments which include: all or a part of a hydrophobic sequence (a sequence above the dashed line); or all or part of a hydrophilic fragment (a sequence below the dashed line). Other fragments include a cysteine residue or as N-glycosylation site.

Figure 12 shows an analysis of the 25278 sulfatase amino acid sequence: αβturn and coil regions; hydrophilicity; amphipathic regions; flexible regions; antigenic index; and surface probability plot.

Figure 13 shows an analysis of the 25278 sulfatase open reading frame for amino acids corresponding to specific functional sites. For the N-glycosylation sites, the actual modified residue is the first amino acid. For cAMP- and cGMP-dependent protein kinase phosphorylation sites, the actual modified residue is the last amino acid. For protein kinase C phosphorylation sites, the actual modified residue is the first amino acid. For casein kinase II phosphorylation sites, the actual modified residue is the first amino acid. For the tyrosine kinase phosphorylation site, the actual modified residue is the last amino acid residue. For N-myristoylation sites, the actual modified residue is the first amino acid. In addition, amidation sites are found from

about amino acids 312-315 and 541-544, and sulfatase signatures are found from about amino acids 139-148 and 91-103.

Figure 14 shows relative expression of 25278 sulfatase mRNA in normal and cancerous human tissues.

Figure 15 shows the 26212 sulfatase cDNA sequence (SEQ ID NO:8) and the deduced amino acid sequence (SEQ ID NO:7). The 26212 sulfatase coding sequence is set forth in SEQ ID NO:14.

Figure 16 shows a 26212 sulfatase hydrophobicity plot. Relative hydrophobic residues are shown above the dashed horizontal line, and relative hydrophilic residues are below the dashed horizontal line. The cysteine residues (cys) and N glycosylation site (Ngly) are indicated by short vertical lines just below the hydropathy trace. The numbers corresponding to the amino acid sequence (shown in SEQ ID NO:7) of 26212 sulfatase are indicated. Polypeptides of the invention include fragments which include: all or a part of a hydrophobic sequence (a sequence above the dashed line); or all or part of a hydrophilic fragment (a sequence below the dashed line). Other fragments include a cysteine residue or as N-glycosylation site.

Figure 17 shows an analysis of the 26212 sulfatase amino acid sequence: αβturn and coil regions; hydrophilicity; amphipathic regions; flexible regions; antigenic index; and surface probability plot.

Figure 18 shows an analysis of the 26212 sulfatase open reading frame for amino acids corresponding to specific functional sites. For the N-glycosylation sites, the actual modified residue is the first amino acid. For cAMP- and cGMP-dependent protein kinase phosphorylation sites, the actual modified residue is the last amino acid. For protein kinase C phosphorylation sites, the actual modified residue is the first amino acid. For casein kinase II phosphorylation sites, the actual modified residue is the first amino acid. For the tyrosine kinase phosphorylation site, the actual modified residue is the last amino acid residue. For N-myristoylation sites, the actual

modified residue is the first amino acid. In addition, sulfatase signature sites are found from about amino acids 168-177 and 120-132.

Figure 19 depicts an alignment of the 22438 sulfatase domain with a consensus amino acid sequence derived from a hidden Markov model. The upper sequence is the consensus amino acid sequence (SEQ ID NO:9), while the lower amino acid sequence corresponds to amino acids 36 to 462 of SEQ ID NO:1.

Figure 20 depicts an alignment of the 23553 sulfatase domain with a consensus amino acid sequence derived from a hidden Markov model. The upper sequence is the consensus amino acid sequence (SEQ ID NO:9), while the lower amino acid sequence corresponds to amino acids 43 to 467 of SEQ ID NO:3.

Figure 21 shows the expression of 23553 in the following human carcinoma cell lines: breast cancer cell lines MCF-7, ZR75, T47D, MDA231, and MDA435; colon cancer cell lines DLD-1, SW480, SW620, HCT116, HT29, and Colo205; lung cancer cell lines NCIH125, NCIH69, NCIH322, NCIH460, and A549. Expression levels were determined by reverse transcriptase(RT) quantitative PCR (Taqman® brand quantitative PCR kit, Applied Biosystems). The quantitative PCR reactions were performed according to the kit manufacturer's instructions.

Figure 22 shows the expression of 23553 in clinical samples of normal human breast tissue and the following human breast tumor tissues: ductal in situ carcinoma (DCIS), invasive ductal carcinoma (IDC), and invasive lobular carcinoma (ILC). Expression levels were determined as described in the description of Figure 21.

Figure 23 shows the expression of 23553 in human clinical samples of normal colon, colon tumor; metastatic liver, and normal liver tissue. Expression levels were determined as described in the description of Figure 21.

Figure 24 shows the expression of 23553 in normal human lung and adenocarcinoma (AC) and squamous cell carcinoma (SCC) lung tumor tissue. Expression levels were determined as described in the description of Figure 21.

Figure 25 shows the expression of 23553 in the following normal human tissues: prostate (column 1), liver (columns 2 and 3), breast (columns 4 and 5), skeletal muscle (column 6), brain (columns 7 and 8), colon (columns 9 and 10), heart (columns 11 and 12), ovary (columns 13 and 14), kidney (columns 15 and 16), lung (columns 17 and 18), vein (columns 19 and 20), trachea (column 21), adipose (columns 22 and 23), small intestine (column 24), thyroid (columns 25 and 26), skin (columns 27 and 28), testes (column 29), placenta (column 30), fetal liver (columns 31 and 32), fetal heart (columns 33 and 34), osteoblasts (undifferentiated, column 35 and primary culture, column 36), fetal spinal cord (column 38), cervix (column 39), spleen (column 40), spinal cord (column 41), thymus (column 42), tonsil (column 43), lymph node (column 44), and aorta (column 45). 23553 was expressed at high levels in trachea, vein, osteoblast, kidney, and testes tissue; significant expression of 23553 was noted in adipose, colon, skeletal muscle, thyroid, and prostate tissues. Expression levels were determined as described in the description of Figure 21.

Figure 26 shows the expression of 23553 in the following human tissues: normal brain (column 1), glioblastoma (columns 2-5), normal breast (column 6), breast tumor (columns 7-9), normal colon (column 10), colon tumor (columns 11-13), normal liver (column 14), metastatic colon (columns 15 and 16), normal lung (column 17), lung tumor (columns 18-20), placenta (column 21), fetal adrenal gland (column 22), normal skin (columns 23 and 24), and adipose (column 25). 23553 was detectable in all tissues tested, with evidence of increased expression levels in breast, colon, and lung tumors. In addition, 23553 was expressed at an elevated level in glioblastoma tissue, as compared to normal brain tissue. Expression levels were determined as described in the description of Figure 21.

Figure 27 depicts an alignment of the 25278 sulfatase domain with a consensus amino acid sequence derived from a hidden Markov model. The upper

sequence is the consensus amino acid sequence (SEQ ID NO:9), while the lower amino acid sequence corresponds to amino acids 47 to 471 of SEQ ID NO:5.

Figure 28 shows the relative expression of 25278 in various human tissues, as follows. Row 1, NDR 19, breast, DCIS (ductal in situ carcinoma); Row 2, MDA 138, breast, normal; Row 3, NDR 01, breast, IDC (invasive ductal carcinoma); Row 4, NDR 15, breast, DC (ductal carcinoma); Row 5, NDR 133, breast, ILC (invasive lobular carcinoma); Row 6, MDA 161, breast, IDC; Row 7, MDA 155, breast, IDC/DCIS; Row 8, PIT 270, lung, normal; Row 9, CHT 427, lung, normal; Row 10, PIT 241, lung, normal; Row 11, PIT 298, lung, normal; Row 12, CHT 800, lung, AC (adenocarcinoma); Row 13, CHT 335, lung, SCC (squamous cell carcinoma); Row 14, CHT447, lung, AC; Row 15, CHT 752, lung, AC; Row 16, CHT 799, lung, AC; Row 17, CHT 369, lung, SCC; Row 18, CHT 369, lung, SCC; Row 19, CHT 371, colon, normal; Row 20, CHT 396, colon, normal; Row 21, CHT 398, colon, normal; Row 22, NDR 104, colon, normal; Row 23, CHT 520, colon, adenocarcinoma; Row 24, CHT 122, colon, adenocarcinoma; Row 25, CHT 536, colon, adenocarcinoma; Row 26, CHT 528, colon, adenocarcinoma; Row 27, CHT 386, colon, adenocarcinoma; Row 28, CHT 372, colon, adenocarcinoma; Row 29, CHT 532, colon, adenocarcinoma; Row 30, CHT 77, liver, metastatic; Row 31, CHT 321, liver, metastatic; Row 32, CHT 84, liver, metastatic; Row 33, NDR 100, liver, metastatic; Row 34, NDR 154, liver, normal; Row 35, CHT 322, liver, normal; Row 36, PIT 51, liver, normal; Row 37, CHT 339, liver, normal; Row 38, PIT 265, breast, normal; Row 39, MDA 335, breast, normal; Row 40, NDR 132, breast, DCIS; Row 41, NDR 13, breast, normal; Row 42, NDR 56, breast, normal.

Figure 29 depicts an alignment of the 26212 sulfatase domain with a consensus amino acid sequence derived from a hidden Markov model. The upper sequence is the consensus amino acid sequence (SEQ ID NO:10), while the lower amino acid sequence corresponds to amino acids 76 to 502 of SEQ ID NO:7.

Figure 30 shows the expression of 26212 in various human endothelial cells, as follows. Proliferating human umbilical vein endothelial cells (HUVEC) (column 1);

arresting HUVEC (column 2); HUVEC minus growth factor (column 3); proliferating cardiac human microvascular endothelial cells (HMVEC) (columns 4 and 6); arresting cardiac HMVEC (columns 5 and 7); proliferating lung HMVEC (columns 8, 11, and 13); arresting lung HMVEC (columns 9, 12, and 14); and lung HMVEC minus growth factor (columns 10 and 15); HEK 293 (non-endothelial) cells (column 16). In six of six independent experiments, 26212 is up-regulated in proliferating endothelial cells as compared to arrested endothelial cells. Further, 26212 expression levels are higher in proliferating endothelial cells than in HEK 293 (non-endothelial) cells. Expression levels were determined as described in the description of Figure 21.

Figure 31 shows the expression of 26212 in the following human tissues. Figure 31A: normal breast (columns 1 and 2), breast tumor (columns 3-9), normal ovary (columns 10 and 11), ovary tumor (columns 12-19), normal lung (columns 20-23), lung tumor (columns 24-31). Figure 31B: normal colon (columns 1-4), colon tumor (columns 5-12), liver metastases (columns 13-16), normal liver (columns 17-18), normal brain (columns19-20), astrocyte (column 21), brain tumor (columns 22-25), arresting human microvascular endothelial cells (column 26), proliferating human microvascular endothelial cells (column 27), placenta (column 28), fetal adrenal tissue (columns 29-30), and fetal liver (column 31). Expression levels were determined as described in the description of Figure 21.

Figure 32 shows 26212 expression in normal human clinical breast samples (columns 1 and 2) and human clinical breast tumor samples (columns 3-9). Expression levels were determined as described in the description of Figure 21.

Figure 33 shows 26212 expression in normal human clinical lung samples (columns 1-4) and human clinical lung tumor samples (columns 5-12). Expression levels were determined as described in the description of Figure 21.

Figure 34 shows the temporal expression of 26212 in human normal and breast cancer epithelial cell lines (MCF10A and MCF3B, respectively) after treatment with epidermal growth factor (EGF). MCF10A cells are shown 0, 0.5, 1, 2, 4, and 8 hours

after treatment with EGF (columns 1-6, respectively). Similarly, MCF3B cells are shown 0, 0.5, 1, 2, 4, and 8 hours after treatment with EGF (columns 7-12, respectively). 26212 is up-regulated in both cell lines. Expression levels were determined as described in the description of Figure 21.

Figure 35 shows expression of 26212 in human hemangiomas and other angiogenic tissues: hemangioma (ONC 101; column 1); hemangioma (ONC 102; column 2); hemangioma (ONC 103; column 3); skin (NDR 295; column 4); fetal heart (BWH4; column 5); normal heart (MPI 849; column 6); spinal cord (CKN 746; column 7); uterine adenocarcinoma (CHT 1424; column 8); and endometrial polyps (CLN 944; column 9). Expression levels were determined as described in the description of Figure 21.

Figure 36 shows expression of 26212 in the following human tissues: normal artery (column 1), normal vein (column 2), aortic smooth muscle cells (SMC), early (column 3), coronary SMC (column 4), static human umbilical vein endothelial cells (HUVEC) (column 5), shear HUVEC (column 6), normal heart (column 7), heart, congestive heart failure (CHF) (column 8), kidney (column 9), skeletal muscle (column 10), normal adipose (column 11), pancreas (column 12), primary osteoblasts (column 13), osteoclasts, differentiated (column 14), normal skin (column 15), normal spinal cord (column 16), normal brain cortex (column 17), normal brain hypothalamus (column 18), nerve (column 19), dorsal root ganglion (DRG) (column 20), glial cells (astrocytes) (column 21), glioblastoma (column 22), normal breast (column 23), breast tumor (column 24), normal ovary (column 25), ovary tumor (column 26), normal prostate (column 27), prostate tumor (column 28), prostate epithelial cells (column 29), normal colon (column 30), colon tumor (column 31), normal lung (column 32), lung tumor (column 33), lung, chronic obstructive pulmonary disease (COPD) (column 34), colon, inflammatory bowel disease (IBD) (column 35), normal liver (column 36), liver fibrosis (column 37), dermal cells, fibroblasts (column 38), normal spleen (column 39), normal tonsil (column 40), lymph node (column 41), small intestine (column 42), skin, decubitus (column 43), synovium (column 44), bone marrow mononuclear cells (BM-MNC) (column 45), and activated peripheral blood mononuclear cells (PBMC) (column

46). The expression levels of 26212 are higher in endothelial and glial cells than in other tissues and cells. Expression levels were determined as described in the description of Figure 21.

DETAILED DESCRIPTION OF THE INVENTION

Sulfatase Polypeptides

The invention is based on the identification of the novel human 22438 sulfatase. *In situ* hybridization experiments showed that this sulfatase is expressed in the following monkey tissues: sub-populations of DRG neurons (mainly in small and medium sized neurons), in spinal cord (interneurons and motor neurons), and in the brain. The sulfatase is also expressed in human brain. The sulfatase cDNA was identified based on consensus motifs or protein domains characteristic of sulfatases and, in particular, arylsulfatase. BLAST analysis has shown homology with human arylsulfatase E, a human iduronate-2-sulfatase, human N-acetylgalactosamine-6-sulfatase, murine arylsulfatase A, and human arylsulfatase A. However, some homology has also been found with other arylsulfatases from various mammalian species, including, but not limited to, human arylsulfatase D, E, F, and B.

The invention is also based on the identification of the novel human 23553 sulfatase. Taqman analysis has shown positive differential expression in breast and colon cancer and in colonic metastases to the liver (Figure 9). This sulfatase has been identified as a glucosamine-6-sulfatase based on ProDom matches and BLAST analysis. Some homology has also been found to human arylsulfatase A, human N-acetylglucosamine-6-sulfatase, and human iduronate-2-sulfatase.

The invention is also based on the identification of the novel human 25278 sulfatase. The sulfatase is differentially expressed in human colon cancer and in colonic metastases to the liver, as determined by Taqman analysis. This sulfatase has been identified as a N-acetylgalactosamine-4-sulfatase by ProDom matching and BLAST homology alignment. Further, based on BLAST analysis, some homology has also been shown to arylsulfatase B and arylsulfatase A.

The invention is also based on the identification of the novel human 26212 sulfatase. This sulfatase has been identified as an arylsulfatase by ProDom matching

and BLAST sequence alignment. Homology has been shown to arylsulfatase B. Some homology has also been found with arylsulfatase F, E, D, and A, as well as with iduronate 2 sulfatase. Arylsulfatase B is also known as N-acetylgalactosamine-4-sulfatase.

Specifically, newly-identified human genes, termed 22438, 23553, 25278, and 26212 sulfatases are provided. These sequences, and other nucleotide sequences encoding the sulfatase proteins or fragments and variants thereof, are referred to as "22438, 23553, 25278, and 26212 sulfatase sequences."

Plasmids containing the sulfatase cDNA inserts were deposited with the Patent Depository of the American Type Culture Collection (ATCC), 10801 University Boulevard, Manassas, Virginia, on ___, April 5, 2000, May 9, 2000, or _____, and assigned Patent Deposit Numbers _____, PTA-1639, PTA-1846, or _____, respectively. The deposits will be maintained under the terms of the Budapest Treaty on the International Recognition of the Deposit of Microorganisms for the Purposes of Patent Procedure. The deposits were made merely as a convenience for those of skill in the art and is not an admission that a deposit is required under 35 U.S.C. §112.

The sulfatase cDNA was identified in human cDNA libraries. Specifically, expressed sequence tags (EST) found in human cDNA libraries, were selected based on homology to known sulfatase sequences. Based on such EST sequences, primers were designed to identify a full length clone from a human cDNA library. Positive clones were sequenced and the overlapping fragments were assembled. The 22438, 23553, 25278, and 26212 sulfatase amino acid sequences are shown in Figures 1, 5, 10, and 15, respectively, and SEQ ID NOS:1, 3, 5, and 7. The 22438, 23553, 25278, and 26212 sulfatase cDNA sequences are shown in Figures 1, 5, 10, and 15 and SEQ ID NOS:2, 4, 6, and 8.

Analysis of the assembled sequences revealed that the cloned cDNA molecules encoded sulfatase-like polypeptides. BLAST analysis indicated that the 23553 sulfatase is a glucosamine-6-sulfatase, that the 25278 sulfatase is an N-acetylgalactosamine-4-sulfatase, that the 22438 is an arylsulfatase with highest homology to arylsulfatase A and E genes and that the 26212 sulfatase is an arylsulfatase with highest homology to the arylsulfatase B gene (N-acetylgalactosamine-4-sulfatase).

The sulfatase sequences of the invention belong to the sulfatase family of molecules having conserved functional features. The term "family" when referring to the proteins and nucleic acid molecules of the invention is intended to mean two or more proteins or nucleic acid molecules having sufficient amino acid or nucleotide sequence identity as defined herein to provide a specific function. Such family members can be naturally-occurring and can be from either the same or different species. For example, a family can contain a first protein of murine origin and an ortholog of that protein of human origin, as well as a second, distinct protein of human origin and a murine ortholog of that protein.

The 22438 sulfatase gene encodes an approximately 2175 nucleotide mRNA transcript having the corresponding cDNA set forth in SEQ ID NO:2. This transcript has an open reading frame which encodes a 525 amino acid protein (SEQ ID NO:1).

The 23553 sulfatase gene encodes an approximately 4321 nucleotide mRNA transcript having the corresponding cDNA set forth in SEQ ID NO:4. This transcript has an open reading frame which encodes an 871 amino acid protein (SEQ ID NO:3).

The 25278 sulfatase gene encodes an approximately 2940 nucleotide mRNA transcript having the corresponding cDNA set forth in SEQ ID NO:6. This transcript has an open reading frame which encodes a 569 amino acid protein (SEQ ID NO:5).

The 26212 sulfatase gene encodes an approximately 2253 nucleotide mRNA transcript having the corresponding cDNA set forth in SEQ ID NO:8. This transcript has an open reading frame which encodes a 599 amino acid protein (SEQ ID NO:7).

Prosite program analysis was used to predict various sites within the 22438 sulfatase protein as shown in Figure 4.

Prosite program analysis was used to predict various sites within the 23553 sulfatase protein as shown in Figure 8.

Prosite program analysis was used to predict various sites within the 25278 sulfatase protein as shown in Figure 13.

Prosite program analysis was used to predict various sites within the 26212 sulfatase protein as shown in Figure 18.

In situ hybridization experiments showed that 22438 is expressed in subpopulations of DRG neurons, spinal cord, and brain, as disclosed hereinabove.

Expression of the 22438 sulfatase mRNA in the above cells and tissues indicates that the sulfatase is likely to be involved in the proper function of and in disorders involving these tissues. Accordingly, the disclosed invention further relates to methods and compositions for the study, modulation, diagnosis and treatment of sulfatase related disorders, especially disorders of these tissues that include, but are not limited to those disclosed herein.

The 23553 sulfatase is differentially expressed in breast and colon cancer and in colonic metastases to the liver. Accordingly, the disclosed invention further relates to methods and compositions for the study, modulation, diagnosis and treatment in these tissues (normal and tumor).

The 25278 sulfatase is differentially expressed in colon tumors and colonic metastases to the liver. Accordingly, the disclosed invention further relates to methods and compositions for the study, modulation, diagnosis and treatment in these normal and tumor tissues.

The 26212 sulfatase is differentially expressed in colon metastases and lung tumors. Accordingly, the disclosed invention further relates to methods and compositions for the study, modulation, diagnosis and treatment in these normal and tumor tissues.

The compositions include sulfatase polypeptides, nucleic acids, vectors, transformed cells and related variants and fragments thereof, as well as agents that modulate expression of the polypeptides and polynucleotides. In particular, the invention relates to the modulation, diagnosis and treatment of sulfatase related disorders as described herein.

Treatment is defined as the application or administration of a therapeutic agent to a patient, or application or administration of a therapeutic agent to an isolated tissue or cell line from a patient, who has a disease, a symptom of disease or a predisposition toward a disease, with the purpose to cure, heal, alleviate, relieve, alter, remedy, ameliorate, improve or affect the disease, the symptoms of disease or the predisposition toward disease. "Subject, as used herein, can refer to a mammal, *e.g.* a human, or to an experimental or animal or disease model. The subject can also be a non-human animal, *e.g.* a horse, cow, goat, or other domestic animal. A therapeutic

agent includes, but is not limited to, small molecules, peptides, antibodies, ribozymes and antisense oligonucleotides.

Disorders involving the brain include, but are not limited to, disorders involving neurons, and disorders involving glia, such as astrocytes, oligodendrocytes, ependymal cells, and microglia; cerebral edema, raised intracranial pressure and herniation, and hydrocephalus; malformations and developmental diseases, such as neural tube defects, forebrain anomalies, posterior fossa anomalies, and syringomyelia and hydromyelia; perinatal brain injury; cerebrovascular diseases, such as those related to hypoxia, ischemia, and infarction, including hypotension, hypoperfusion, and low-flow states--global cerebral ischemia and focal cerebral ischemia--infarction from obstruction of local blood supply, intracranial hemorrhage, including intracerebral (intraparenchymal) hemorrhage, subarachnoid hemorrhage and ruptured berry aneurysms, and vascular malformations, hypertensive cerebrovascular disease, including lacunar infarcts, slit hemorrhages, and hypertensive encephalopathy; infections, such as acute meningitis, including acute pyogenic (bacterial) meningitis and acute aseptic (viral) meningitis, acute focal suppurative infections, including brain abscess, subdural empyema, and extradural abscess, chronic bacterial meningoencephalitis, including tuberculosis and mycobacterioses, neurosyphilis, and neuroborreliosis (Lyme disease), viral meningoencephalitis, including arthropod-borne (Arbo) viral encephalitis, *Herpes simplex* virus Type 1, *Herpes simplex* virus Type 2, *Varicalla-zoster* virus (*Herpes zoster*), cytomegalovirus, poliomyelitis, rabies, and human immunodeficiency virus 1, including HIV-1 meningoencephalitis (subacute encephalitis), vacuolar myelopathy, AIDS-associated myopathy, peripheral neuropathy, and AIDS in children, progressive multifocal leukoencephalopathy, subacute sclerosing panencephalitis, fungal meningoencephalitis, other infectious diseases of the nervous system; transmissible spongiform encephalopathies (prion diseases); demyelinating diseases, including multiple sclerosis, multiple sclerosis variants, acute disseminated encephalomyelitis and acute necrotizing hemorrhagic encephalomyelitis, and other diseases with demyelination; degenerative diseases, such as degenerative diseases affecting the cerebral cortex, including Alzheimer disease and Pick disease, degenerative diseases of basal ganglia and brain stem, including Parkinsonism, idiopathic Parkinson disease (paralysis agitans), progressive

supranuclear palsy, corticobasal degeneration, multiple system atrophy, including striatonigral degeneration, Shy-Drager syndrome, and olivopontocerebellar atrophy, and Huntington disease; spinocerebellar degenerations, including spinocerebellar ataxias, including Friedreich ataxia, and ataxia-telanglectasia, degenerative diseases affecting motor neurons, including amyotrophic lateral sclerosis (motor neuron disease), bulbospinal atrophy (Kennedy syndrome), and spinal muscular atrophy; inborn errors of metabolism, such as leukodystrophies, including Krabbe disease, metachromatic leukodystrophy, adrenoleukodystrophy, Pelizaeus-Merzbacher disease, and Canavan disease, mitochondrial encephalomyopathies, including Leigh disease and other mitochondrial encephalomyopathies; toxic and acquired metabolic diseases, including vitamin deficiencies such as thiamine (vitamin B_1) deficiency and vitamin B_{12} deficiency, neurologic sequelae of metabolic disturbances, including hypoglycemia, hyperglycemia, and hepatic encephatopathy, toxic disorders, including carbon monoxide, methanol, ethanol, and radiation, including combined methotrexate and radiation-induced injury; tumors, such as gliomas, including astrocytoma, including fibrillary (diffuse) astrocytoma and glioblastoma multiforme, pilocytic astrocytoma, pleomorphic xanthoastrocytoma, and brain stem glioma, oligodendroglioma, and ependymoma and related paraventricular mass lesions, neuronal tumors, poorly differentiated neoplasms, including medulloblastoma, other parenchymal tumors, including primary brain lymphoma, germ cell tumors, and pineal parenchymal tumors, meningiomas, metastatic tumors, paraneoplastic syndromes, peripheral nerve sheath tumors, including schwannoma, neurofibroma, and malignant peripheral nerve sheath tumor (malignant schwannoma), and neurocutaneous syndromes (phakomatoses), including neurofibromotosis, including Type 1 neurofibromatosis (NF1) and TYPE 2 neurofibromatosis (NF2), tuberous sclerosis, and Von Hippel-Lindau disease.

Furthermore, as disclosed in the background hereinabove, specific disorders have been associated with function of the various sulfatases. Accordingly, the sulfatases disclosed herein, having homology to specific sulfatases as disclosed herein, are useful for diagnosis and treatment of the disorders associated with sulfatase dysfunction as disclosed herein and to modulation of gene expression in the affected tissues.

The sequences of the invention find use in diagnosis of disorders involving an increase or decrease in sulfatase expression relative to normal expression, such as a proliferative disorder, a differentiative disorder, or a developmental disorder. The sequences also find use in modulating sulfatase-related responses. By "modulating" is intended the upregulating or downregulating of a response. That is, the compositions of the invention affect the targeted activity in either a positive or negative fashion.

The invention relates to novel sulfatases, having the deduced amino acid sequence shown in Figures 1, 5, 10, and 15 (SEQ ID NOS:1, 3, 5, and 7) or having the amino acid sequences encoded by the deposited cDNAs, Patent Deposit Numbers _____, PTA-1639, PTA-1846, or _____. The deposited sequences, as well as the polypeptides encoded by the sequences, are incorporated herein by reference and control in the event of any conflict, such as a sequencing error, with description in this application.

Thus, the present invention provides an isolated or purified sulfatase polypeptides and variants and fragments thereof. "Sulfatase polypeptide" or "sulfatase protein" refers to the polypeptide in SEQ ID NOS:1, 3, 5, or 7 or encoded by the deposited cDNAs. The term "sulfatase protein" or "sulfatase polypeptide," however, further includes the numerous variants described herein, as well as fragments derived from the full-length sulfatase and variants.

Sulfatase polypeptides can be purified to homogeneity. It is understood, however, that preparations in which the polypeptide is not purified to homogeneity are useful and considered to contain an isolated form of the polypeptide. The critical feature is that the preparation allows for the desired function of the polypeptide, even in the presence of considerable amounts of other components. Thus, the invention encompasses various degrees of purity.

As used herein, a polypeptide is said to be "isolated" or "purified" when it is substantially free of cellular material when it is isolated from recombinant and non-recombinant cells, or free of chemical precursors or other chemicals when it is chemically synthesized. A polypeptide, however, can be joined to another polypeptide with which it is not normally associated in a cell and still be considered "isolated" or "purified."

In one embodiment, the language "substantially free of cellular material" includes preparations of sulfatase having less than about 30% (by dry weight) other

proteins (i.e., contaminating protein), less than about 20% other proteins, less than about 10% other proteins, or less than about 5% other proteins. When the polypeptide is recombinantly produced, it can also be substantially free of culture medium, i.e., culture medium represents less than about 20%, less than about 10%, or less than about 5% of the volume of the protein preparation.

The sulfatase polypeptide is also considered to be isolated when it is part of a membrane preparation or is purified and then reconstituted with membrane vesicles or liposomes.

The language "substantially free of chemical precursors or other chemicals" includes preparations of the sulfatase polypeptide in which it is separated from chemical precursors or other chemicals that are involved in its synthesis. The language "substantially free of chemical precursors or other chemicals" includes, but is not limited to, preparations of the polypeptide having less than about 30% (by dry weight) chemical precursors or other chemicals, less than about 20% chemical precursors or other chemicals, less than about 10% chemical precursors or other chemicals, or less than about 5% chemical precursors or other chemicals.

In one embodiment, the sulfatase polypeptide comprises the amino acid sequence shown in SEQ ID NOS:1, 3, 5, or 7. However, the invention also encompasses sequence variants. By "variants" is intended proteins or polypeptides having an amino acid sequence that is at least about 45%, 55%, 65%, preferably about 75%, 85%, 95%, or 98% identical to the amino acid sequence of SEQ ID NOS:1,3,5, or 7. Variants also include polypeptides encoded by the cDNA insert of the plasmid deposited with ATCC as Patent Deposit Numbers _____, PTA-1639, PTA-1846, or _____, or polypeptides encoded by a nucleic acid molecule that hybridizes to the nucleic acid molecule of SEQ ID NOS:2, 4, 6, 8, 11, 12, 13, or 14, or a complement thereof, under stringent conditions. In another embodiment, a variant of an isolated polypeptide of the present invention differs, by at least 1, but less than 5, 10, 20, 50, or 100 amino acid residues from the sequence shown in SEQ ID NO:1, 3, 5, or 7. If alignment is needed for this comparison the sequences should be aligned for maximum identity. "Looped" out sequences from deletions or insertions, or mismatches, are considered differences. Such variants generally retain the functional activity of the 22438-like, 23553-like, 25278-like, or 26212-like proteins of the invention. Variants include

polypeptides that differ in amino acid sequence due to natural allelic variation or mutagenesis.

Variants include a substantially homologous protein encoded by the same genetic locus in an organism, i.e., an allelic variant. Variants also encompass proteins derived from other genetic loci in an organism, but having substantial homology to the sulfatase of SEQ ID NOS:1, 3, 5, or 7. Variants also include proteins substantially homologous to the sulfatase but derived from another organism, i.e., an ortholog. Variants also include proteins that are substantially homologous to the sulfatase that are produced by chemical synthesis. Variants also include proteins that are substantially homologous to the sulfatase that are produced by recombinant methods. Variants retain the biological activity (for example, sulfatase activity) of the polypeptide set forth by the reference sequence (SEQ ID NOS: 1, 3, 5, or 7). It is understood, however, that variants exclude any amino acid sequences disclosed prior to the invention.

Preferred sulfatase polypeptides of the present invention have an amino acid sequence sufficiently identical to the amino acid sequence of SEQ ID NOS:1, 3, 5, or 7. The term "sufficiently identical" is used herein to refer to a first amino acid or nucleotide sequence that contains a sufficient or minimum number of identical or equivalent (e.g., with a similar side chain) amino acid residues or nucleotides to a second amino acid or nucleotide sequence such that the first and second amino acid or nucleotide sequences have a common structural domain and/or common functional activity. For example, amino acid or nucleotide sequences that contain a common structural domain having at least about 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98% or 99% identity are defined herein as sufficiently identical.

In one embodiment, a variant of the 23553 sulfatase is greater than 92% homologous. In another embodiment, a variant of the 25278 sulfatase is greater than 50% identical. In another embodiment, the 26212 sulfatase is greater than 50% identical.

To determine the percent identity of two amino acid sequences, or of two nucleic acid sequences, the sequences are aligned for optimal comparison purposes (e.g., gaps can be introduced in one or both of a first and a second amino acid or nucleic acid sequence for optimal alignment and non-homologous sequences can be disregarded for comparison purposes). In a preferred embodiment, the length of a

reference sequence aligned for comparison purposes is at least 30%, preferably at least 40%, more preferably at least 50%, even more preferably at least 60%, and even more preferably at least 70%, 80%, 90%, 100% of the length of the reference sequence. The amino acid residues or nucleotides at corresponding amino acid positions or nucleotide positions are then compared. When a position in the first sequence is occupied by the same amino acid residue or nucleotide as the corresponding position in the second sequence, then the molecules are identical at that position (as used herein amino acid or nucleic acid "identity" is equivalent to amino acid or nucleic acid "homology"). The percent identity between the two sequences is a function of the number of identical positions shared by the sequences, taking into account the number of gaps, and the length of each gap, which need to be introduced for optimal alignment of the two sequences.

The comparison of sequences and determination of percent identity between two sequences can be accomplished using a mathematical algorithm. In a preferred embodiment, the percent identity between two amino acid sequences is determined using the Needleman and Wunsch (1970) *J. Mol. Biol. 48*:444-453 algorithm which has been incorporated into the GAP program in the GCG software package (available at http://www.gcg.com), using either a Blossum 62 matrix or a PAM250 matrix, and a gap weight of 16, 14, 12, 10, 8, 6, or 4 and a length weight of 1, 2, 3, 4, 5, or 6. In yet another preferred embodiment, the percent identity between two nucleotide sequences is determined using the GAP program in the GCG software package (available at http://www.gcg.com), using a NWSgapdna.CMP matrix and a gap weight of 40, 50, 60, 70, or 80 and a length weight of 1, 2, 3, 4, 5, or 6. A particularly preferred set of parameters (and the one that should be used if the practitioner is uncertain about what parameters should be applied to determine if a molecule is within a sequence identity or homology limitation of the invention) is using a Blossum 62 scoring matrix with a gap open penalty of 12, a gap extend penalty of 4, and a frameshift gap penalty of 5.

The percent identity between two amino acid or nucleotide sequences can be determined using the algorithm of E. Meyers and W. Miller (1989) *CABIOS 4*:11-17 which has been incorporated into the ALIGN program (version 2.0), using a PAM120 weight residue table, a gap length penalty of 12 and a gap penalty of 4.

The nucleic acid and protein sequences described herein can be used as a "query sequence" to perform a search against public databases to, for example, identify other family members or related sequences. Such searches can be performed using the NBLAST and XBLAST programs (version 2.0) of Altschul, *et al.* (1990) *J. Mol. Biol. 215*:403-10. BLAST nucleotide searches can be performed with the NBLAST program, score = 100, wordlength = 12 to obtain nucleotide sequences homologous to the nucleic acid molecules of the invention. BLAST protein searches can be performed with the XBLAST program, score = 50, wordlength = 3 to obtain amino acid sequences homologous to the protein molecules of the invention. To obtain gapped alignments for comparison purposes, Gapped BLAST can be utilized as described in Altschul *et al.* (1997) *Nucleic Acids Res. 25(17)*:3389-3402. When utilizing BLAST and Gapped BLAST programs, the default parameters of the respective programs (e.g., XBLAST and NBLAST) can be used. See http://www.ncbi.nlm.nih.gov.

The invention also encompasses polypeptides having a lower degree of identity but having sufficient similarity so as to perform one or more of the same functions performed by the sulfatase. Similarity is determined by conservative amino acid substitution, as shown in Table 1. Such substitutions are those that substitute a given amino acid in a polypeptide by another amino acid of like characteristics. Conservative substitutions are likely to be phenotypically silent. Typically seen as conservative substitutions are the replacements, one for another, among the aliphatic amino acids Ala, Val, Leu, and Ile; interchange of the hydroxyl residues Ser and Thr, exchange of the acidic residues Asp and Glu, substitution between the amide residues Asn and Gln, exchange of the basic residues Lys and Arg and replacements among the aromatic residues Phe, Tyr. Guidance concerning which amino acid changes are likely to be phenotypically silent are found in Bowie *et al., Science 247*:1306-1310 (1990).

TABLE 1. Conservative Amino Acid Substitutions.

Aromatic	Phenylalanine
	Tryptophan
	Tyrosine
Hydrophobic	Leucine
	Isoleucine
	Valine
Polar	Glutamine
	Asparagine
Basic	Arginine
	Lysine
	Histidine
Acidic	Aspartic Acid
	Glutamic Acid
Small	Alanine
	Serine
	Threonine
	Methionine
	Glycine

A variant polypeptide can differ in amino acid sequence by one or more substitutions, deletions, insertions, inversions, fusions, and truncations or a combination of any of these. Variant polypeptides can be fully functional or can lack function in one or more activities. Thus, in the present case, variations can affect the function, for example, of one or more of regions including a metal (e.g., Ca^{++})-binding domain, activation domain, sulfatase catalytic domain, the region containing a propeptide, regulatory regions, substrate binding regions, regions involved in membrane association or subcellular localization, regions involved in post-

translational modification, for example, by phosphorylation, and regions that are important for effector function (i.e., agents that act upon the protein, such as in the conversion of cysteine to 2-amino-3-oxoproprionic acid or serine semi-aldehyde).

Fully functional variants typically contain only conservative variation or variation in non-critical residues or in non-critical regions. Functional variants can also contain substitution of similar amino acids, which results in no change or an insignificant change in function. Alternatively, such substitutions may positively or negatively affect function to some degree.

Non-functional variants typically contain one or more non-conservative amino acid substitutions, deletions, insertions, inversions, or truncation or a substitution, insertion, inversion, or deletion in a critical residue or critical region.

As indicated, variants can be naturally-occurring or can be made by recombinant means or chemical synthesis to provide useful and novel characteristics for the sulfatase polypeptide. This includes preventing immunogenicity from pharmaceutical formulations by preventing protein aggregation.

Useful variations further include alteration of functional activity. For example, one embodiment involves a variation at the substrate binding site that results in binding but not hydrolysis or more or less hydrolysis of the substrate than wild type. A further useful variation at the same site can result in altered affinity for the substrate. Useful variations also include changes that provide for affinity for another substrate. Useful variations further include the ability to bind an effector molecule with greater or lesser affinity, such as not to bind or to bind but not release it. Further useful variations include alteration in the ability of the propeptide to be cleaved by a cleavage protein, including alteration in the binding or recognition site. Further, the cleavage site can also be modified so that recognition and cleavage are by a different protease. A specific useful variation involves a variation in the ability to be bound or activated by the enzyme that activates the sulfatase by the conversion of cysteine to 2-3-oxoproprionic acid or serine semi-aldehyde. Further variation could include a variation in the specificity of metal binding.

Another useful variation provides a fusion protein in which one or more domains or subregions are operationally fused to one or more domains, subregions, or motifs from another sulfatase. For example, a transmembrane domain from a protein can be

introduced into the sulfatase such that the protein is anchored in the cell surface. Other permutations include changing the number of sulfatase domains, and mixing of sulfatase domains from different sulfatase families, so that substrate specificity is altered. Mixing these various domains can allow the formation of novel sulfatase molecules with different host cell, subcellular localization, substrate, and effector molecule (one that acts on the sulfatase) specificity.

The term "substrate" is intended to refer not only to the sulfated substrate that is cleaved by the sulfatase domain, but to refer to any component with which the polypeptide interacts in order to produce an effect on that component or a subsequent biological effect that is a result of interacting with that component. This can include, but is not limited to, for example, interaction with the sulfatase activation enzyme and components involved in the conversion of 3′ phosphoadenosine 5′ phosphosulfate to adenosine 3′ 5′ biphosphate.

Amino acids that are essential for function can be identified by methods known in the art, such as site-directed mutagenesis or alanine-scanning mutagenesis (Cunningham *et al.* (1985) *Science 244*:1081-1085). The latter procedure introduces single alanine mutations at every residue in the molecule. The resulting mutant molecules are then tested for biological activity, such as peptide bond hydrolysis *in vitro* or related biological activity, such as proliferative activity. Sites that are critical for binding can also be determined by structural analysis such as crystallization, nuclear magnetic resonance or photoaffinity labeling (Smith *et al.* (1992) *J. Mol. Biol. 224*:899-904; de Vos *et al.* (1992) *Science 255*:306-312).

The invention thus also includes polypeptide fragments of the sulfatases. Fragments can be derived from the amino acid sequence shown in SEQ ID NOS:1, 3, 5, or 7. However, the invention also encompasses fragments of the variants of the sulfatase polypeptides as described herein. The fragments to which the invention pertains, however, are not to be construed as encompassing fragments that may be disclosed prior to the present invention.

A fragment can comprise at least about 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 30, 35, 40, 45, 50 or more contiguous amino acids. Fragments can retain one or more of the biological activities of the protein, for example as discussed

above, as well as fragments that can be used as an immunogen to generate sulfatase antibodies.

For example, for the 25278 sulfatase, the invention encompasses amino acid fragments greater than 5 amino acids, particularly from regions up to around nucleotide 450 and beyond around nucleotide 1520. Specific fragments which may be excluded include those that are underlined in Figure 1. However, even in regions between around nucleotide 450 to around nucleotide 1520, fragments include those that are five or greater excluding those which may have been disclosed prior to the present invention.

For the 23553 sulfatase, fragments particularly include fragments of 5 amino acids or more up to around nucleotide 670.

For the 26212 sulfatase, for example, fragments containing 5 or more amino acids up to about nucleotide 572 are particularly encompassed by the invention. However, fragments of 5 amino acids or more encoded by around nucleotide 572 to around nucleotide 1985 are also encompassed by the invention with the understanding that such fragments do not encompass those which may have been disclosed prior to the invention. For example, these can include the sections underlined in Figure 15.

Biologically active fragments (peptides which are, for example, about 5, 10, 15, 20, 25, 30, 35, 40, 50, 100 or more amino acids in length) can comprise a functional site. Such sites include but are not limited to those discussed above, such as a catalytic site, regulatory site, site important for substrate recognition or binding, regions containing a sulfatase domain or motif, phosphorylation sites, glycosylation sites, and other functional sites disclosed herein.

Fragments, for example, can extend in one or both directions from the functional site to encompass 5, 10, 15, 20, 30, 40, 50, or up to 100 amino acids. Further, fragments can include sub-fragments of the specific sites or regions disclosed herein, which sub-fragments retain the function of the site or region from which they are derived.

The invention also provides fragments with immunogenic properties. These contain an epitope-bearing portion of the sulfatase polypeptide and variants. These epitope-bearing peptides are useful to raise antibodies that bind specifically to a sulfatase polypeptide or region or fragment. These peptides can contain at least 10, 12, at least 14, or between at least about 15 to about 30 amino acids. The epitope-bearing sulfatase

polypeptides may be produced by any conventional means (Houghten, R.A. (1985) *Proc. Natl. Acad. Sci. USA 82*:5131-5135). Simultaneous multiple peptide synthesis is described in U.S. Patent No. 4,631,211.

Non-limiting examples of antigenic polypeptides that can be used to generate antibodies include but are not limited to peptides derived from extracellular regions. Regions having a high antigenicity index are shown in Figures 3, 7, 12, and 17. However, intracellularly-made antibodies ("intrabodies") are also encompassed, which would recognize intracellular peptide regions.

Fragments can be discrete (not fused to other amino acids or polypeptides) or can be within a larger polypeptide. Further, several fragments can be comprised within a single larger polypeptide. In one embodiment a fragment designed for expression in a host can have heterologous pre- and pro-polypeptide regions fused to the amino terminus of the sulfatase polypeptide fragment and an additional region fused to the carboxyl terminus of the fragment.

The invention thus provides chimeric or fusion proteins. These comprise a sulfatase peptide sequence operatively linked to a heterologous peptide having an amino acid sequence not substantially homologous to the sulfatase polypeptide. "Operatively linked" indicates that the sulfatase polypeptide and the heterologous peptide are fused in-frame. The heterologous peptide can be fused to the N-terminus or C-terminus of the sulfatase polypeptide or can be internally located.

In one embodiment the fusion protein does not affect sulfatase function *per se*. For example, the fusion protein can be a GST-fusion protein in which sulfatase sequences are fused to the N- or C-terminus of the GST sequences. Other types of fusion proteins include, but are not limited to, enzymatic fusion proteins, for example beta-galactosidase fusions, yeast two-hybrid GAL4 fusions, poly-His fusions and Ig fusions. Such fusion proteins, particularly poly-His fusions, can facilitate the purification of recombinant sulfatase polypeptide. In certain host cells (e.g., mammalian host cells), expression and/or secretion of a protein can be increased by using a heterologous signal sequence. Therefore, in another embodiment, the fusion protein contains a heterologous signal sequence at its C- or N-terminus.

EP-A-O 464 533 discloses fusion proteins comprising various portions of immunoglobulin constant regions. The Fc is useful in therapy and diagnosis and thus

results, for example, in improved pharmacokinetic properties (EP-A 0232 262). In drug discovery, for example, human proteins have been fused with Fc portions for the purpose of high-throughput screening assays to identify antagonists (Bennett *et al.* (1995) *J. Mol. Recog. 8*:52-58 (1995) and Johanson *et al. J. Biol. Chem. 270*:9459-9471). Thus, this invention also encompasses soluble fusion proteins containing a sulfatase polypeptide and various portions of the constant regions of heavy or light chains of immunoglobulins of various subclass (IgG, IgM, IgA, IgE). Preferred as immunoglobulin is the constant part of the heavy chain of human IgG, particularly IgG1, where fusion takes place at the hinge region. For some uses it is desirable to remove the Fc after the fusion protein has been used for its intended purpose, for example when the fusion protein is to be used as antigen for immunizations. In a particular embodiment, the Fc part can be removed in a simple way by a cleavage sequence, which is also incorporated and can be cleaved with factor Xa.

A chimeric or fusion protein can be produced by standard recombinant DNA techniques. For example, DNA fragments coding for the different protein sequences are ligated together in-frame in accordance with conventional techniques. In another embodiment, the fusion gene can be synthesized by conventional techniques including automated DNA synthesizers. Alternatively, PCR amplification of gene fragments can be carried out using anchor primers which give rise to complementary overhangs between two consecutive gene fragments which can subsequently be annealed and re-amplified to generate a chimeric gene sequence (see Ausubel *et al.* (1992) *Current Protocols in Molecular Biology*). Moreover, many expression vectors are commercially available that already encode a fusion moiety (e.g., a GST protein). A sulfatase-encoding nucleic acid can be cloned into such an expression vector such that the fusion moiety is linked in-frame to sulfatase.

Another form of fusion protein is one that directly affects sulfatase functions. Accordingly, a sulfatase polypeptide is encompassed by the present invention in which one or more of the sulfatase regions (or parts thereof) has been replaced by heterologous or homologous regions (or parts thereof) from another sulfatase. Accordingly, various permutations are possible, for example, as discussed above. Thus, chimeric sulfatases can be formed in which one or more of the native domains or subregions has been

duplicated, removed, or replaced by another. This includes but is not limited to catalytic sulfatase or substrate binding domains, and regions involved in activation.

It is understood however that such regions could be derived from a sulfatase that has not yet been characterized. Moreover, sulfatase function can be derived from peptides that contain these functions but are not in a sulfatase family.

The isolated 22438 sulfatase protein can be purified from cells that naturally express it, such as DRG neurons, including small and medium sized neurons, spinal cord, including interneurons and motor neurons, and brain, especially purified from cells that have been altered to express it (recombinant), or synthesized using known protein synthesis methods.

The isolated 23553 sulfatase protein can be purified from cells that naturally express it, such as cells from any of the tissues shown in Figures 9 and 21-26, especially purified from cells that have been altered to express it (recombinant), or synthesized using known protein synthesis methods.

The isolated 25278 sulfatase protein can be purified from cells that naturally express it, such as cells from any of the tissues shown in Figures 14 and 28, especially purified from cells that have been altered to express it (recombinant), or synthesized using known protein synthesis methods.

The isolated 26212 sulfatase protein can be purified from cells that naturally express it, such as cells from any of the tissues shown in Figures 30-36, especially purified from cells that have been altered to express it (recombinant), or synthesized using known protein synthesis methods.

In one embodiment, the protein is produced by recombinant DNA techniques. For example, a nucleic acid molecule encoding the sulfatase polypeptide is cloned into an expression vector, the expression vector introduced into a host cell and the protein expressed in the host cell. The protein can then be isolated from the cells by an appropriate purification scheme using standard protein purification techniques.

Polypeptides often contain amino acids other than the 20 amino acids commonly referred to as the 20 naturally-occurring amino acids. Further, many amino acids, including the terminal amino acids, may be modified by natural processes, such as processing and other post-translational modifications, or by chemical modification techniques well known in the art. Common modifications that occur naturally in

polypeptides are described in basic texts, detailed monographs, and the research literature, and they are well known to those of skill in the art.

Accordingly, the polypeptides also encompass derivatives or analogs in which a substituted amino acid residue is not one encoded by the genetic code, in which a substituent group is included, in which the mature polypeptide is fused with another compound, such as a compound to increase the half-life of the polypeptide (for example, polyethylene glycol), or in which the additional amino acids are fused to the mature polypeptide, such as a leader or secretory sequence or a sequence for purification of the mature polypeptide or a pro-protein sequence.

Known modifications include, but are not limited to, acetylation, acylation, ADP-ribosylation, amidation, covalent attachment of flavin, covalent attachment of a heme moiety, covalent attachment of a nucleotide or nucleotide derivative, covalent attachment of a lipid or lipid derivative, covalent attachment of phosphatidylinositol, cross-linking, cyclization, disulfide bond formation, demethylation, formation of covalent crosslinks, formation of cystine, formation of pyroglutamate, formylation, gamma carboxylation, glycosylation, GPI anchor formation, hydroxylation, iodination, methylation, myristoylation, oxidation, proteolytic processing, phosphorylation, prenylation, racemization, selenoylation, sulfation, transfer-RNA mediated addition of amino acids to proteins such as arginylation, and ubiquitination.

Such modifications are well-known to those of skill in the art and have been described in great detail in the scientific literature. Several particularly common modifications, glycosylation, lipid attachment, sulfation, gamma-carboxylation of glutamic acid residues, hydroxylation and ADP-ribosylation, for instance, are described in most basic texts, such as *Proteins - Structure and Molecular Properties*, 2nd ed., T.E. Creighton, W. H. Freeman and Company, New York (1993). Many detailed reviews are available on this subject, such as by Wold, F., *Posttranslational Covalent Modification of Proteins*, B.C. Johnson, Ed., Academic Press, New York 1-12 (1983); Seifter *et al.* (1990) *Meth. Enzymol. 182*: 626-646) and Rattan *et al.* (1992) *Ann. N.Y. Acad. Sci. 663*:48-62).

As is also well known, polypeptides are not always entirely linear. For instance, polypeptides may be branched as a result of ubiquitination, and they may be circular, with or without branching, generally as a result of post-translation events, including

natural processing events and events brought about by human manipulation which do not occur naturally. Circular, branched and branched circular polypeptides may be synthesized by non-translational natural processes and by synthetic methods.

Modifications can occur anywhere in a polypeptide, including the peptide backbone, the amino acid side-chains and the amino or carboxyl termini. Blockage of the amino or carboxyl group in a polypeptide, or both, by a covalent modification, is common in naturally-occurring and synthetic polypeptides. For instance, the aminoterminal residue of polypeptides made in *E. coli*, prior to proteolytic processing, almost invariably will be N-formylmethionine.

The modifications can be a function of how the protein is made. For recombinant polypeptides, for example, the modifications will be determined by the host cell posttranslational modification capacity and the modification signals in the polypeptide amino acid sequence. Accordingly, when glycosylation is desired, a polypeptide should be expressed in a glycosylating host, generally a eukaryotic cell. Insect cells often carry out the same posttranslational glycosylations as mammalian cells and, for this reason, insect cell expression systems have been developed to efficiently express mammalian proteins having native patterns of glycosylation. Similar considerations apply to other modifications.

The same type of modification may be present in the same or varying degree at several sites in a given polypeptide. Also, a given polypeptide may contain more than one type of modification.

Polypeptide Uses

The protein sequences of the present invention can be used as a "query sequence" to perform a search against public databases to, for example, identify other family members or related sequences. Such searches can be performed using the NBLAST and XBLAST programs (version 2.0) of Altschul *et al.* (1990) *J. Mol. Biol.* 215:403-10. BLAST nucleotide searches can be performed with the NBLAST program, score = 100, wordlength = 12 to obtain nucleotide sequences homologous to the nucleic acid molecules of the invention. BLAST protein searches can be performed with the XBLAST program, score = 50, wordlength = 3 to obtain amino acid sequences homologous to the proteins of the invention. To obtain gapped

alignments for comparison purposes, Gapped BLAST can be utilized as described in Altschul *et al.*, (1997) *Nucleic Acids Res.* 25(17):3389-3402. When utilizing BLAST and Gapped BLAST programs, the default parameters of the respective programs (e.g., XBLAST and NBLAST) can be used. See http://www.ncbi.nlm.nih.gov.

Sulfatase polypeptides are useful for producing antibodies specific for sulfatase, regions, or fragments. Regions having a high antigenicity index score are shown in Figures 3, 7, 12, and 17.

Sulfatase polypeptides are useful for biological assays related to sulfatases. Such assays involve any of the known sulfatase functions or activities or properties useful for diagnosis and treatment of sulfatase-related conditions, including those in the references cited herein, which are incorporated by reference for these assays, functions, and disorders.

These assays include, but are not limited to, binding to and/or cleaving specific substrates to produce fragments, steady state levels of sulfated compounds, cysteine modification, and biological assays related to the functions produced by sulfated compounds. Specific substrates useful for assays related to sulfate conjugate hydrolysis include but are not limited to xenobiotics, thyroid hormones, steroids, and catechols. Specific sulfate conjugates include, but are not limited to, 3α-sulfatolithocholyltaurine, sulfate conjugates of estrone, 4-methylumbelliferone, and harmol, sulfated cartilage and proteoglycans, 4-nitrophenol, simple phenols, hydroxyarylamines, iodothyronines, catecholamines, 1-naphthyl, salbutamol, estrogens, ethinylestradiol, equilenin, diethylstilbestrol, androgens, cholesterol bile salts, pregnenolone, benzylic alcohols, glycolipidsulfates, complex carbohydrates such as dermatan and chondrotin sulfate, steroid sulfate, sulfate conjugates of xenobiotics, cholesterol sulfate, xenobiotic phenyls, *o*-cresol, vanillan, eugenol, *m*-cresol, thymol, ethyl-4,4-dihydroxybenzoate, *p*-cresol, sesamol, methyl-2,6-dihydroxy-4-methylbenzyloate, methyl-2,4-dihydroxybenzoate, methyl-3,5-dihydroxybenzoate, tyramine, dopamine, 5 hydroxytryptamine, pyrogallol, 4-nitrocatecholsulfate, estrone sulfate, metabolites of the cytochrome P450 mono-oxygenase system, dihydroepiandrosterone sulfate (DHEAS), minoxidil, cicletanine, sulfated mutagens and carcinogens, such as aromatic amines (including heterocyclic amines), and benzylic alcohols of chemicals such as polycyclic aromatic hydrocarbons, saffrole and estragole, glycosaminoglycans, sulfolipids, betahydroxysteroids, sulfate

esters of chromogenic or fluorogenic aromatic compounds, cerebroside sulfate, keritan sulfate, and heparan sulfate. Substrates also include any in the references cited herein, which are incorporated herein by reference for these substrates. Accordingly the assays include, but are not limited to, these sulfated substrates and biological effects of sulfation or desulfation of these substrates and associated biochemical, cellular, or phenotypic effects of sulfation of desulfation, and any of the other biological or functional properties of these proteins, including, but not limited to, those disclosed herein, and in any reference cited herein which is incorporated herein by reference for the disclosure of these properties and for the assays based on these properties. Further, assays may relate to changes in the protein, *per se*, and on the effects of these changes, for example, activation of the sulfatase by modification of a cysteine residue as disclosed herein, cleavage of the propeptide by a proteinase, induction of expression of the protein *in vivo*, inhibition of function, as well as any other effects on the protein mentioned herein or cited in any reference herein, which are incorporated herein by reference for these effects and for the subsequent biological consequences of these effects.

Sulfatase polypeptides are also useful in drug screening assays, in cell-based or cell-free systems. Cell-based systems can be native, i.e., cells that normally express sulfatase, such as those discussed above, especially tumor cells, as a biopsy, or expanded in cell culture. In one embodiment, however, cell-based assays involve recombinant host cells expressing sulfatase. Accordingly, these drug-screening assays can be based on effects on protein function as described above for biological assays useful for diagnosis and treatment.

Determining the ability of the test compound to interact with a sulfatase can also comprise determining the ability of the test compound to preferentially bind to the polypeptide as compared to the ability of a known binding molecule to bind to the polypeptide.

The polypeptides can be used to identify compounds that modulate sulfatase activity. Such compounds, for example, can increase or decrease affinity or rate of binding to substrate, compete with substrate for binding to sulfatase, or displace substrate bound to sulfatase. Both sulfatase and appropriate variants and fragments can be used in high-throughput screens to assay candidate compounds for the ability to bind to sulfatase. These compounds can be further screened against a functional sulfatase to

determine the effect of the compound on sulfatase activity. Compounds can be identified that activate (agonist) or inactivate (antagonist) sulfatase to a desired degree. Modulatory methods can be performed *in vitro* (e.g., by culturing the cell with the agent) or, alternatively, *in vivo* (e.g., by administering the agent to a subject).

Sulfatase polypeptides can be used to screen a compound for the ability to stimulate or inhibit interaction between sulfatase protein and a target molecule that normally interacts with the sulfatase, for example, substrate of the sulfatase domain. The assay includes the steps of combining sulfatase protein with a candidate compound under conditions that allow the sulfatase protein or fragment to interact with the target molecule, and to detect the formation of a complex between the sulfatase protein and the target or to detect the biochemical consequence of the interaction with the sulfatase and the target.

Determining the ability of the sulfatase to bind to a target molecule can also be accomplished using a technology such as real-time Bimolecular Interaction Analysis (BIA). Sjolander *et al.* (1991) *Anal. Chem. 63*:2338-2345 and Szabo *et al.* (1995) *Curr Opin. Struct. Biol. 5*:699-705. As used herein, "BIA" is a technology for studying biospecific interactions in real time, without labeling any of the interactants (e.g., BIAcore™). Changes in the optical phenomenon surface plasmon resonance (SPR) can be used as an indication of real-time reactions between biological molecules.

The test compounds of the present invention can be obtained using any of the numerous approaches in combinatorial library methods known in the art, including: biological libraries; spatially addressable parallel solid phase or solution phase libraries; synthetic library methods requiring deconvolution; the 'one-bead one-compound' library method; and synthetic library methods using affinity chromatography selection. The biological library approach is limited to polypeptide libraries, while the other four approaches are applicable to polypeptide, non-peptide oligomer or small molecule libraries of compounds (Lam, K.S. (1997) *Anticancer Drug Des. 12*:145).

Examples of methods for the synthesis of molecular libraries can be found in the art, for example in DeWitt *et al.* (1993) *Proc. Natl. Acad. Sci. USA 90*:6909; Erb *et al.* (1994) *Proc. Natl. Acad. Sci. USA 91*:11422; Zuckermann *et al.* (1994). *J. Med.*

Chem. 37:2678; Cho *et al.* (1993) *Science 261*:1303; Carell *et al.* (1994) *Angew. Chem. Int. Ed. Engl. 33*:2059; Carell *et al.* (1994) *Angew. Chem. Int. Ed. Engl. 33*:2061; and in Gallop *et al.* (1994) *J. Med. Chem. 37*:1233. Libraries of compounds may be presented in solution (e.g., Houghten (1992) *Biotechniques 13*:412-421), or on beads (Lam (1991) *Nature 354*:82-84), chips (Fodor (1993) *Nature 364*:555-556), bacteria (Ladner USP 5,223,409), spores (Ladner USP '409), plasmids (Cull *et al.* (1992) *Proc. Natl. Acad. Sci. USA 89*:1865-1869) or on phage (Scott and Smith (1990) *Science 249*:386-390); (Devlin (1990) *Science 249*:404-406); (Cwirla *et al.* (1990) *Proc. Natl. Acad. Sci. 97*:6378-6382); (Felici (1991) *J. Mol. Biol. 222*:301-310); (Ladner *supra*).

Candidate compounds include, for example, 1) peptides such as soluble peptides, including Ig-tailed fusion peptides and members of random peptide libraries (see, e.g., Lam *et al.* (1991) *Nature 354*:82-84; Houghten *et al.* (1991) *Nature 354*:84-86) and combinatorial chemistry-derived molecular libraries made of D- and/or L- configuration amino acids; 2) phosphopeptides (e.g., members of random and partially degenerate, directed phosphopeptide libraries, see, e.g., Songyang *et al.* (1993) *Cell 72*:767-778); 3) antibodies (e.g., polyclonal, monoclonal, humanized, anti-idiotypic, chimeric, and single chain antibodies as well as Fab, $F(ab')_2$, Fab expression library fragments, and epitope-binding fragments of antibodies); 4) small organic and inorganic molecules (e.g., molecules obtained from combinatorial and natural product libraries); substrate analogs including, but not limited to, substrates disclosed herein.

One candidate compound is a soluble full-length sulfatase or fragment that competes for substrate. Other candidate compounds include mutant sulfatases or appropriate fragments containing mutations that affect sulfatase function and compete for substrate. Accordingly, a fragment that competes for substrate, for example with a higher affinity, or a fragment that binds substrate but does not process or otherwise affect it, is encompassed by the invention.

The invention provides other end points to identify compounds that modulate (stimulate or inhibit) sulfatase activity. The assays typically involve an assay of cellular events that indicate sulfatase activity. Thus, the expression of genes that are up- or down-regulated in response to sulfatase activity can be assayed. In one embodiment, the regulatory region of such genes can be operably linked to a marker that is easily

detectable, such as luciferase. Alternatively, modification of the sulfatase could also be measured.

Any of the biological or biochemical functions mediated by the sulfatase can be used as an endpoint assay. These include any of the biochemical or biochemical/biological events described herein, in any reference cited herein, incorporated by reference for these endpoint assay targets, and other functions known to those of ordinary skill in the art. Specific end points can include, but are not limited to, the events resulting from expression (or lack thereof) of sulfatase activity. With respect to disorders, this would include, but not be limited to, effects on function, differentiation, and proliferation, which can be assayed, as well as the biological effects of function, such as disorders discussed hereinabove and in the references cited hereinabove which are incorporated herein by reference for the disorders disclosed in those references and other disorders and pathology. In the case of the 22438 sulfatase, models of pain can be used as an end point. In the case of the 23553 and 25278 sulfatases, tumor progression can be used as an end point. In the case of the 26212 sulfatase, tumor angiogenesis and/or tumor progression can be used as an end point.

Binding and/or activating compounds can also be screened by using chimeric sulfatase proteins in which one or more regions, segments, sites, and the like, as disclosed herein, or parts thereof, can be replaced by heterologous and homologous counterparts derived from other sulfatases. For example, a catalytic region can be used that interacts with a different substrate specificity and/or affinity than the native sulfatase. Accordingly, a different set of components is available as an end-point assay for activation. As a further alternative, the site of modification by an effector protein, for example, activation or phosphorylation, can be replaced with the site for a different effector protein. Activation can also be detected by a reporter gene containing an easily detectable coding region operably linked to a transcriptional regulatory sequence that is part of the native pathway in which sulfatase is involved.

Sulfatase polypeptides are also useful in competition binding assays in methods designed to discover compounds that interact with the sulfatase. Thus, a compound is exposed to a sulfatase polypeptide under conditions that allow the compound to bind or to otherwise interact with the polypeptide. Soluble sulfatase polypeptide is also added to

the mixture. If the test compound interacts with the soluble sulfatase polypeptide, it decreases the amount of complex formed or activity from the sulfatase target. This type of assay is particularly useful in cases in which compounds are sought that interact with specific regions of the sulfatase. Thus, the soluble polypeptide that competes with the target sulfatase region is designed to contain peptide sequences corresponding to the region of interest.

Another type of competition-binding assay can be used to discover compounds that interact with specific functional sites. As an example, bindable substrate analog and a candidate compound can be added to a sample of the sulfatase. Compounds that interact with the sulfatase at the same site as the substrate or analog will reduce the amount of complex formed between the sulfatase and the substrate or analog. Accordingly, it is possible to discover a compound that specifically prevents interaction between the sulfatase and the component. Another example involves adding a candidate compound to a sample of sulfatase and cleavable substrate. A compound that competes with the substrate will reduce the amount of hydrolysis or binding of the substrate to the sulfatase. Accordingly, compounds can be discovered that directly interact with the sulfatase and compete with the substrate. Such assays can involve any other component that interacts with the sulfatase.

To perform cell free drug screening assays, it is desirable to immobilize either sulfatase, or fragment, or its target molecule to facilitate separation of complexes from uncomplexed forms of one or both of the proteins, as well as to accommodate automation of the assay.

Techniques for immobilizing proteins on matrices can be used in the drug screening assays. In one embodiment, a fusion protein can be provided which adds a domain that allows the protein to be bound to a matrix. For example, glutathione-S-transferase/sulfatase fusion proteins can be adsorbed onto glutathione sepharose beads (Sigma Chemical, St. Louis, MO) or glutathione derivatized microtitre plates, which are then combined with the cell lysates (e.g., ^{35}S-labeled) and the candidate compound, and the mixture incubated under conditions conducive to complex formation (e.g., at physiological conditions for salt and pH). Following incubation, the beads are washed to remove any unbound label, and the matrix immobilized and radiolabel determined directly, or in the supernatant after the complexes is dissociated. Alternatively, the

complexes can be dissociated from the matrix, separated by SDS-PAGE, and the level of sulfatase-binding protein found in the bead fraction quantitated from the gel using standard electrophoretic techniques. For example, either the polypeptide or its target molecule can be immobilized utilizing conjugation of biotin and streptavidin using techniques well known in the art. Alternatively, antibodies reactive with the protein but which do not interfere with binding of the protein to its target molecule can be derivatized to the wells of the plate, and the protein trapped in the wells by antibody conjugation. Preparations of a sulfatase-binding target component, such as substrate or activating enzyme, and a candidate compound are incubated in sulfatase-presenting wells and the amount of complex trapped in the well can be quantitated. Methods for detecting such complexes, in addition to those described above for the GST-immobilized complexes, include immunodetection of complexes using antibodies reactive with the sulfatase target molecule, or which are reactive with the sulfatase and compete with the target molecule; as well as enzyme-linked assays which rely on detecting an enzymatic activity associated with the target molecule.

Modulators of sulfatase activity identified according to these drug screening assays can be used to treat a subject with a disorder related to the sulfatase, by treating cells that express the sulfatase. These methods of treatment include the steps of administering the modulators of sulfatase activity in a pharmaceutical composition as described herein, to a subject in need of such treatment.

The 23553, 25278, and 26212 sulfatases are differentially expressed in tumor cells as disclosed herein. Accordingly, these sulfatases are relevant to these disorders and relevant as well to differentiation, function, and growth of the tissues giving rise to the tumors. The 22438 sulfatase is expressed as described above, and accordingly is relevant for disorders involving these tissues. Disorders include, but are not limited to, those discussed hereinabove. Moreover, since the gene is expressed in the central nervous system, this sulfatase is relevant for the treatment of pain.

Sulfatase polypeptides are thus useful for treating a sulfatase-associated disorder characterized by aberrant expression or activity of a sulfatase. "Aberrant expression" or "misexpression", as used herein, refers to a non-wild type pattern of gene expression, at the RNA or protein level. It includes: expression at non-wild type levels, i.e., over or under expression; a pattern of expression that differs from wild

type in terms of the time or stage at which the gene is expressed, e.g., increased or decreased expression (as compared with wild type) at a predetermined developmental period or stage; a pattern of expression that differs from wild type in terms of decreased expression (as compared with wild type) in a predetermined cell type or tissue type; a pattern of expression that differs from wild type in terms of the splicing size, amino acid sequence, post-transitional modification, or biological activity of the expressed polypeptide; a pattern of expression that differs from wild type in terms of the effect of an environmental stimulus or extracellular stimulus on expression of the gene, e.g., a pattern of increased or decreased expression (as compared with wild type) in the presence of an increase or decrease in the strength of the stimulus.

In one embodiment, the method involves administering an agent (e.g., an agent identified by a screening assay described herein), or combination of agents that modulates (e.g., upregulates or downregulates) expression or activity of the protein. In another embodiment, the method involves administering sulfatase as therapy to compensate for reduced or aberrant expression or activity of the protein.

Methods for treatment include but are not limited to the use of soluble sulfatase or fragments of sulfatase protein that compete for substrate or any other component that directly interacts with sulfatase, or any of the enzymes that modify the sulfatase. These sulfatases or fragments can have a higher affinity for the target so as to provide effective competition.

Stimulation of activity is desirable in situations in which the protein is abnormally downregulated and/or in which increased activity is likely to have a beneficial effect. Likewise, inhibition of activity is desirable in situations in which the protein is abnormally upregulated and/or in which decreased activity is likely to have a beneficial effect. In one example of such a situation, a subject has a disorder characterized by aberrant development or cellular differentiation. In another example, the subject has a disorder characterized by an aberrant hematopoietic response. In another example, it is desirable to achieve tissue regeneration in a subject.

In yet another aspect of the invention, the proteins of the invention can be used as "bait proteins" in a two-hybrid assay or three-hybrid assay (see, e.g., U.S. Patent No. 5,283,317; Zervos *et al.* (1993) *Cell 72*:223-232; Madura *et al.* (1993) *J. Biol. Chem. 268*:12046-12054; Bartel *et al.* (1993) *Biotechniques 14*:920-924; Iwabuchi *et*

al. (1993) *Oncogene 8*:1693-1696; and Brent WO 94/10300), to identify other proteins (captured proteins) which bind to or interact with the proteins of the invention and modulate their activity.

Sulfatase polypeptides also are useful to provide a target for diagnosing a disease or predisposition to disease mediated by the sulfatase, including, but not limited to, those diseases disclosed herein, in the references cited herein, and as disclosed above in the background. Accordingly, methods are provided for detecting the presence, or levels of the sulfatase in a cell, tissue, or organism. The method involves contacting a biological sample with a compound capable of interacting with the sulfatase such that the interaction can be detected. One agent for detecting a sulfatase is an antibody capable of selectively binding to the sulfatase. A biological sample includes tissues, cells and biological fluids isolated from a subject, as well as tissues, cells and fluids present within a subject.

The sulfatase also provides a target for diagnosing active disease, or predisposition to disease, in a patient having a variant sulfatase. Thus, sulfatase can be isolated from a biological sample and assayed for the presence of a genetic mutation that results in an aberrant protein. This includes amino acid substitution, deletion, insertion, rearrangement, (as the result of aberrant splicing events), and inappropriate post-translational modification. Analytic methods include altered electrophoretic mobility, altered tryptic peptide digest, altered sulfatase activity in cell-based or cell-free assays, such as by alteration in substrate binding or degradation, or ability to be activated by the activation enzyme, or antibody-binding pattern, altered isoelectric point, direct amino acid sequencing, and any other of the known assay techniques useful for detecting mutations in a protein in general or in a sulfatase specifically, such as are disclosed herein.

In vitro techniques for detection of sulfatase include enzyme linked immunosorbent assays (ELISAs), Western blots, immunoprecipitations and immunofluorescence. Alternatively, the protein can be detected *in vivo* in a subject by introducing into the subject a labeled anti-sulfatase antibody. For example, the antibody can be labeled with a radioactive marker whose presence and location in a subject can be detected by standard imaging techniques. Particularly useful are methods, which detect

the allelic variant of sulfatase expressed in a subject, and methods, which detect fragments of sulfatase in a sample.

Sulfatase polypeptides are also useful in pharmacogenomic analysis. Pharmacogenomics deal with clinically significant hereditary variations in the response to drugs due to altered drug disposition and abnormal action in affected persons. See, e.g., Eichelbaum, M. (1996) *Clin. Exp. Pharmacol. Physiol. 23(10-11)*:983-985, and Linder, M.W. (1997) *Clin. Chem. 43(2)*:254-266. The clinical outcomes of these variations result in severe toxicity of therapeutic drugs in certain individuals or therapeutic failure of drugs in certain individuals as a result of individual variation in metabolism. Thus, the genotype of the individual can determine the way a therapeutic compound acts on the body or the way the body metabolizes the compound. Further, the activity of drug metabolizing enzymes affects both the intensity and duration of drug action. Thus, the pharmacogenomics of the individual permit the selection of effective compounds and effective dosages of such compounds for prophylactic or therapeutic treatment based on the individual's genotype. The discovery of genetic polymorphisms in some drug metabolizing enzymes has explained why some patients do not obtain the expected drug effects, show an exaggerated drug effect, or experience serious toxicity from standard drug dosages. Polymorphisms can be expressed in the phenotype of the extensive metabolizer and the phenotype of the poor metabolizer. Accordingly, genetic polymorphism may lead to allelic protein variants of sulfatase in which one or more of sulfatase functions in one population is different from those in another population. The polypeptides thus allow a target to ascertain a genetic predisposition that can affect treatment modality. Thus, in a peptide-based treatment, polymorphism may give rise to catalytic regions that are more or less active. Accordingly, dosage would necessarily be modified to maximize the therapeutic effect within a given population containing the polymorphism. As an alternative to genotyping, specific polymorphic polypeptides could be identified.

Sulfatase polypeptides are also useful for monitoring therapeutic effects during clinical trials and other treatment. Thus, the therapeutic effectiveness of an agent that is designed to increase or decrease gene expression, protein levels or sulfatase activity can be monitored over the course of treatment using sulfatase polypeptides as an end-point target. The monitoring can be, for example, as follows: (i) obtaining a pre-

administration sample from a subject prior to administration of the agent; (ii) detecting the level of expression or activity of the protein in the pre-administration sample; (iii) obtaining one or more post-administration samples from the subject; (iv) detecting the level of expression or activity of the protein in the post-administration samples; (v) comparing the level of expression or activity of the protein in the pre-administration sample with the protein in the post-administration sample or samples; and (vi) increasing or decreasing the administration of the agent to the subject accordingly.

Antibodies

The invention also provides antibodies that selectively bind to the sulfatase and its variants and fragments. An antibody is considered to selectively bind, even if it also binds to other proteins that are not substantially homologous with the sulfatase. These other proteins share homology with a fragment or domain of sulfatase. This conservation in specific regions gives rise to antibodies that bind to both proteins by virtue of the homologous sequence. In this case, it would be understood that antibody binding to the sulfatase is still selective.

Antibodies can be polyclonal or monoclonal. An intact antibody, or a fragment thereof (e.g. Fab or $F(ab')_2$) can be used. An appropriate immunogenic preparation can be derived from native, recombinantly expressed, or chemically synthesized peptides.

To generate antibodies, an isolated sulfatase polypeptide is used as an immunogen to generate antibodies using standard techniques for polyclonal and monoclonal antibody preparation. Either the full-length protein or antigenic peptide fragment can be used. Regions having a high antigenicity index are disclosed hereinabove.

Antibodies are preferably prepared from these regions or from discrete fragments in these regions. However, antibodies can be prepared from any region of the peptide as described herein. A preferred fragment produces an antibody that diminishes or completely prevents substrate hydrolysis or binding. Antibodies can be developed against the entire sulfatase or domains of the sulfatase as described herein, for example, the substrate binding region, sulfatase motif, or subregions thereof.

Antibodies can also be developed against other specific functional sites as disclosed herein.

The antigenic peptide can comprise a contiguous sequence of at least 12, 14, 15-20, 20-25, or 25-30 or more amino acid residues. In one embodiment, fragments correspond to regions that are located on the surface of the protein, e.g., hydrophilic regions. These fragments are not to be construed, however, as encompassing any fragments, which may be disclosed prior to the invention.

Detection can be facilitated by coupling (i.e., physically linking) the antibody to a detectable substance. Examples of detectable substances include various enzymes, prosthetic groups, fluorescent materials, luminescent materials, bioluminescent materials, and radioactive materials. Examples of suitable enzymes include horseradish peroxidase, alkaline phosphatase, β-galactosidase, or acetylcholinesterase; examples of suitable prosthetic group complexes include streptavidin/biotin and avidin/biotin; examples of suitable fluorescent materials include umbelliferone, fluorescein, fluorescein isothiocyanate, rhodamine, dichlorotriazinylamine fluorescein, dansyl chloride or phycoerythrin; an example of a luminescent material includes luminol; examples of bioluminescent materials include luciferase, luciferin, and aequorin, and examples of suitable radioactive material include ^{125}I, ^{131}I, ^{35}S or ^{3}H.

Antibody Uses

The antibodies can be used to isolate a sulfatase by standard techniques, such as affinity chromatography or immunoprecipitation. The antibodies can facilitate the purification of the natural sulfatase from cells and recombinantly produced sulfatase expressed in host cells.

The antibodies are useful to detect the presence of a sulfatase in cells or tissues to determine the pattern of expression of the sulfatase among various tissues in an organism and over the course of normal development. The antibodies can be used to detect a sulfatase *in situ*, *in vitro*, or in a cell lysate or supernatant in order to evaluate the abundance and pattern of expression. Antibody detection of circulating fragments of the full length sulfatase can be used to identify sulfatase turnover. In addition, the antibodies can be used to assess abnormal tissue distribution or abnormal expression during development.

Further, the antibodies can be used to assess sulfatase expression in disease states such as in active stages of the disease or in an individual with a predisposition toward disease related to sulfatase function. When a disorder is caused by an inappropriate tissue distribution, developmental expression, or level of expression of sulfatase protein, the antibody can be prepared against the normal sulfatase protein. If a disorder is characterized by a specific mutation in sulfatase, antibodies specific for this mutant protein can be used to assay for the presence of the specific mutant sulfatase. However, intracellularly-made antibodies ("intrabodies") are also encompassed, which would recognize intracellular sulfatase peptide regions.

The antibodies can also be used to assess normal and aberrant subcellular localization of cells in the various tissues in an organism. Antibodies can be developed against the whole sulfatase or portions of the sulfatase.

The diagnostic uses can be applied, not only in genetic testing, but also in monitoring a treatment modality. Accordingly, where treatment is ultimately aimed at correcting sulfatase expression level or the presence of aberrant sulfatases and aberrant tissue distribution or developmental expression, antibodies directed against the sulfatase or relevant fragments can be used to monitor therapeutic efficacy.

Additionally, antibodies are useful in pharmacogenomic analysis. Thus, antibodies prepared against polymorphic sulfatase can be used to identify individuals that require modified treatment modalities.

The antibodies are also useful as diagnostic tools as an immunological marker for aberrant sulfatase analyzed by electrophoretic mobility, isoelectric point, tryptic peptide digest, and other physical assays known to those in the art.

The antibodies are also useful for tissue typing. Thus, where a specific sulfatase has been correlated with expression in a specific tissue, antibodies that are specific for this sulfatase can be used to identify a tissue type.

The antibodies are also useful in forensic identification. Accordingly, where an individual has been correlated with a specific genetic polymorphism resulting in a specific polymorphic protein, an antibody specific for the polymorphic protein can be used as an aid in identification.

The antibodies are also useful for inhibiting sulfatase function, for example, substrate binding, or sulfatase activity. For example, sulfatase activity may be measured

by the ability to form a binding complex with a sulfated conjugate, such as disclosed herein.

These uses can also be applied in a therapeutic context in which treatment involves inhibiting sulfatase function. An antibody can be used, for example, to block substrate binding. Antibodies can be prepared against specific fragments containing sites required for function or against intact sulfatase associated with a cell.

Completely human antibodies are particularly desirable for therapeutic treatment of human patients. For an overview of this technology for producing human antibodies, see Lonberg *et al.* (1995) *Int. Rev. Immunol. 13*:65-93. For a detailed discussion of this technology for producing human antibodies and human monoclonal antibodies and protocols for producing such antibodies, e.g., U.S. Patent 5,625,126; U.S. Patent 5,633,425; U.S. Patent 5,569,825; U.S. Patent 5,661,016; and U.S. Patent 5,545,806.

The invention also encompasses kits for using antibodies to detect the presence of a sulfatase protein in a biological sample. The kit can comprise antibodies such as a labeled or labelable antibody and a compound or agent for detecting the sulfatase in a biological sample; means for determining the amount of sulfatase in the sample; and means for comparing the amount of sulfatase in the sample with a standard. The compound or agent can be packaged in a suitable container. The kit can further comprise instructions for using the kit to detect the sulfatase.

Polynucleotides

The nucleotide sequences in SEQ ID NOS:2, 4, 6, and 8 were obtained by sequencing the deposited human cDNAs. Accordingly, the sequences of the deposited clones are controlling as to any discrepancies between the two and any reference to a sequence of SEQ ID NOS:2, 4, 6, or 8, includes reference to the sequence of the deposited cDNA.

The specifically disclosed cDNA comprises the coding region and 5' and 3' untranslated sequences in SEQ ID NOS:2, 4, 6, or 8. The coding sequences of the cDNA's are set forth in SEQ ID NOS:11, 12, 13, and 14.

The invention provides isolated polynucleotides encoding the novel sulfatases. The term "sulfatase polynucleotide" or "sulfatase nucleic acid" refers to the sequences

shown in SEQ ID NOS:2, 4, 6, 8, 11, 12, 13, or 14, or in the deposited cDNAs. The term "sulfatase polynucleotide" or "sulfatase nucleic acid" further includes variants and fragments of sulfatase polynucleotides.

Generally, nucleotide sequence variants of the invention will have at least 60%, 65%, 70%, 75%, 80%, 85%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, or 99% identity to one of the nucleotide sequences disclosed herein.

An "isolated" sulfatase nucleic acid is one that is separated from other nucleic acid present in the natural source of sulfatase nucleic acid. Preferably, an "isolated" nucleic acid is free of sequences which naturally flank sulfatase nucleic acid (i.e., sequences located at the 5' and 3' ends of the nucleic acid) in the genomic DNA of the organism from which the nucleic acid is derived. However, there can be some flanking nucleotide sequences, for example up to about 5KB. The important point is that the sulfatase nucleic acid is isolated from flanking sequences such that it can be subjected to the specific manipulations described herein, such as recombinant expression, preparation of probes and primers, and other uses specific to the sulfatase nucleic acid sequences. In one embodiment, the sulfatase nucleic acid comprises only the coding region.

Moreover, an "isolated" nucleic acid molecule, such as a cDNA or RNA molecule, can be substantially free of other cellular material, or culture medium when produced by recombinant techniques, or chemical precursors or other chemicals when chemically synthesized. However, the nucleic acid molecule can be fused to other coding or regulatory sequences and still be considered isolated.

In some instances, the isolated material will form part of a composition (for example, a crude extract containing other substances), buffer system or reagent mix. In other circumstances, the material may be purified to essential homogeneity, for example as determined by PAGE or column chromatography such as HPLC. Preferably, an isolated nucleic acid comprises at least about 50, 80 or 90% (on a molar basis) of all macromolecular species present.

For example, recombinant DNA molecules contained in a vector are considered isolated. Further examples of isolated DNA molecules include recombinant DNA molecules maintained in heterologous host cells or purified (partially or substantially) DNA molecules in solution. Isolated RNA molecules include *in vivo* or *in vitro* RNA transcripts of the isolated DNA molecules of the present invention. Isolated nucleic acid

molecules according to the present invention further include such molecules produced synthetically.

In some instances, the isolated material will form part of a composition (or example, a crude extract containing other substances), buffer system or reagent mix. In other circumstances, the material may be purified to essential homogeneity, for example as determined by PAGE or column chromatography such as HPLC. Preferably, an isolated nucleic acid comprises at least about 50, 80 or 90% (on a molar basis) of all macromolecular species present.

Sulfatase polynucleotides can encode the mature protein plus additional amino or carboxyterminal amino acids, or amino acids interior to the mature polypeptide (when the mature form has more than one polypeptide chain, for instance). Such sequences may play a role in processing of a protein from precursor to a mature form, facilitate protein trafficking, prolong or shorten protein half-life or facilitate manipulation of a protein for assay or production, among other things. As generally is the case *in situ*, the additional amino acids may be processed away from the mature protein by cellular enzymes.

Sulfatase polynucleotides include, but are not limited to, the sequence encoding the mature polypeptide alone, the sequence encoding the mature polypeptide and additional coding sequences, such as a leader or secretory sequence (e.g., a pre-pro or pro-protein sequence), the sequence encoding the mature polypeptide, with or without the additional coding sequences, plus additional non-coding sequences, for example introns and non-coding 5' and 3' sequences such as transcribed but non-translated sequences that play a role in transcription, mRNA processing (including splicing and polyadenylation signals), ribosome binding and stability of mRNA. In addition, the polynucleotide may be fused to a marker sequence encoding, for example, a peptide that facilitates purification.

Sulfatase polynucleotides can be in the form of RNA, such as mRNA, or in the form DNA, including cDNA and genomic DNA obtained by cloning or produced by chemical synthetic techniques or by a combination thereof. The nucleic acid, especially DNA, can be double-stranded or single-stranded. Single-stranded nucleic acid can be the coding strand (sense strand) or the non-coding strand (anti-sense strand).

The invention further provides variant sulfatase polynucleotides, and fragments thereof, that differ from the nucleotide sequence shown in SEQ ID NOS:2, 4, 6, 8, 11, 12, 13, or 14 due to degeneracy of the genetic code and thus encode the same protein as that encoded by a nucleotide sequence shown in SEQ ID NOS:2, 4, 6, 8, 11, 12, 13, or 14.

Alternatively, a nucleic acid molecule that is a fragment of a 22438-like nucleotide sequence of the present invention comprises a nucleotide sequence consisting of nucleotides 1-100, 100-200, 200-300, 300-400, 400-500, 500-600, 600-700, 700-900, 900-1000, 1000-1100, 1100-1200, 1200-1300, 1300-1400, 1400-1500, 1500-1600, 1600-1700, 1700-1800, 1800-1900, 1900-2000, 2000-2100, 2100-2175 of SEQ ID NO:2.

A nucleic acid molecule that is a fragment of a 23553-like nucleotide sequence of the present invention comprises a nucleotide sequence consisting of nucleotides 1-100, 100-200, 200-300, 300-400, 400-500, 500-600, 600-700, 700-900, 900-1000, 1000-1100, 1100-1200, 1200-1300, 1300-1400, 1400-1500, 1500-1600, 1600-1700, 1700-1800, 1800-1900, 1900-2000, 2000-2100, 2100-2200, 2200-2300, 2300-2400, 2400-2500, 2500-2600, 2600-2700, 2700-2800, 2800-2900, 2900-3000, 3000-3100, 3100-3200, 3200-3300, 3300-3400, 3400-3500, 3500-3600, 3600-3700, 3700-3800, 3800-3900, 3900-4000, 4000-4100, 4100-4200, 4200-4300, 4300-4321 of SEQ ID NO:4.

A nucleic acid molecule that is a fragment of a 25278-like nucleotide sequence of the present invention comprises a nucleotide sequence consisting of nucleotides 1-100, 100-200, 200-300, 300-400, 400-500, 500-600, 600-700, 700-900, 900-1000, 1000-1100, 1100-1200, 1200-1300, 1300-1400, 1400-1500, 1500-1600, 1600-1700, 1700-1800, 1800-1900, 1900-2000, 2000-2100, 2100-2200, 2200-2300, 2300-2400, 2400-2500, 2500-2600, 2600-2700, 2700-2800, 2800-2900, 2900-2940 of SEQ ID NO:6.

A nucleic acid molecule that is a fragment of a 26212-like nucleotide sequence of the present invention comprises a nucleotide sequence consisting of nucleotides 1-100, 100-200, 200-300, 300-400, 400-500, 500-600, 600-700, 700-900, 900-1000, 1000-1100, 1100-1200, 1200-1300, 1300-1400, 1400-1500, 1500-1600, 1600-1700, 1700-1800, 1800-1900, 1900-2000, 2000-2100, 2100-2200, 2200-2253 of SEQ ID NO:8.

The invention also provides sulfatase nucleic acid molecules encoding the variant polypeptides described herein. Such polynucleotides may be naturally occurring, such as allelic variants (same locus), homologs (different locus), and orthologs (different organism), or may be constructed by recombinant DNA methods or by chemical synthesis. Such non-naturally occurring variants may be made by mutagenesis techniques, including those applied to polynucleotides, cells, or organisms. Accordingly, as discussed above, the variants can contain nucleotide substitutions, deletions, inversions and insertions.

Typically, variants have a substantial identity with a nucleic acid molecules of SEQ ID NOS:2, 4, 6, 8, 11, 12, 13, or 14, and the complements thereof. Variation can occur in either or both the coding and non-coding regions. The variations can produce both conservative and non-conservative amino acid substitutions.

Orthologs, homologs, and allelic variants can be identified using methods well known in the art. These variants comprise a nucleotide sequence encoding a sulfatase that is typically at least about 40-45%, 45-50%, 50-55%, 55-60%, 60-65%, 65-70%, 70-75%, more typically at least about 75-80% or 80-85%, and most typically at least about 85-90% or 90-95% or more homologous to the nucleotide sequence shown in SEQ ID NOS:2, 4, 6 or 8, or a fragment of this sequence. Such nucleic acid molecules can readily be identified as being able to hybridize under stringent conditions, to the nucleotide sequence shown in SEQ ID NOS:2, 4, 6, 8, 11, 12, 13, or 14, or a fragment of the sequence.

In the case of the 23553 sulfatase, in one embodiment, a variant is greater than 65% homologous with respect to nucleotide sequence. For the 25278 sulfatase, in one embodiment, a variant is greater than 50-60% homologous with respect to nucleotide sequence. With respect to the 26212 sulfatase, in one embodiment, a variant is greater than about 65-75% homologous with respect to nucleotide sequence.

It is understood that stringent hybridization does not indicate substantial homology where it is due to general homology, such as polyA$^+$ sequences, or sequences common to all or most proteins, sulfatases, arylsulfatases, glucosamine-6-sulfatases, N-acetylgalactosamine-4-sulfatases, or any of the sulfatases to which the sulfatases of the present invention have shown homology by BLAST analysis, for example, regions to arylsulfatases A, B, C, D, E, F, IDS, and the like. Moreover, it is understood that

variants do not include any of the nucleic acid sequences that may have been disclosed prior to the invention.

As used herein, the term "hybridizes under stringent conditions" describes conditions for hybridization and washing. Stringent conditions are known to those skilled in the art and can be found in *Current Protocols in Molecular Biology* John Wiley & Sons, N.Y. (1989), 6.3.1-6.3.6. Aqueous and nonaqueous methods are described in that reference and either can be used. A preferred, example of stringent hybridization conditions are hybridization in 6X sodium chloride/sodium citrate (SSC) at about 45°C, followed by one or more washes in 0.2X SSC, 0.1% SDS at 50°C. Another example of stringent hybridization conditions are hybridization in 6X sodium chloride/sodium citrate (SSC) at about 45°C, followed by one or more washes in 0.2X SSC, 0.1% SDS at 55°C. A further example of stringent hybridization conditions are hybridization in 6X sodium chloride/sodium citrate (SSC) at about 45°C, followed by one or more washes in 0.2X SSC, 0.1% SDS at 60°C. Preferably, stringent hybridization conditions are hybridization in 6X sodium chloride/sodium citrate (SSC) at about 45°C, followed by one or more washes in 0.2X SSC, 0.1% SDS at 65°C. Particularly preferred stringency conditions (and the conditions that should be used if the practitioner is uncertain about what conditions should be applied to determine if a molecule is within a hybridization limitation of the invention) are 0.5M Sodium Phosphate, 7% SDS at 65°C, followed by one or more washes at 0.2X SSC, 1% SDS at 65°C. Preferably, an isolated nucleic acid molecule of the invention that hybridizes under stringent conditions to the sequence of SEQ ID NOS:2, 4, 6, 8, 11, 12, 13, or 14 corresponds to a naturally-occurring nucleic acid molecule. As used herein, a "naturally-occurring" nucleic acid molecule refers to an RNA or DNA molecule having a nucleotide sequence that occurs in nature (e.g., encodes a natural protein).

The present invention also provides isolated nucleic acids that contain a single or double stranded fragment or portion that hybridizes under stringent conditions to the nucleotide sequence of SEQ ID NOS:2, 4, 6, 8, 11, 12, 13, or 14, or the complements of SEQ ID NOS:2, 4, 6, 8, 11, 12, 13, or 14. In one embodiment, the nucleic acid consists of a portion of a nucleotide sequence of SEQ ID NOS:2, 4, 6, 8, 11, 12, 13, or 14 and the complements. The nucleic acid fragments of the invention

are at least about 10-15, preferably at least about 15-20 or 20-25 contiguous nucleotides, and can be 30, 33, 35, 40, 50, 60, 70, 75, 80, 90, 100, 200, 500 or more nucleotides in length. Longer fragments, for example, 600 or more nucleotides in length, which encode antigenic proteins or polypeptides described herein are also useful.

In the case of the 23553 sulfatase, in one embodiment, fragments are derived from nucleotide 1 to about nucleotide 670 and comprise 5-10 and 10-20 contiguous base pairs, and particularly greater than 18. For this sulfatase, in another embodiment, a fragment is derived from around nucleotide 3008 to 3514 and comprises around 5-10 and 10-20 contiguous nucleotides. In other embodiments for this sulfatase, a fragment is derived from around nucleotide 3994 to 4321 and is about 5-10 or 10-20 contiguous nucleotides. For the 25278, in one embodiment, a fragment is derived from around nucleotide 130 to around nucleotide 454 and comprises a contiguous sequence of about 5-10 or 10-20 nucleotides. In another embodiment, the fragment is derived from around nucleotide 454 to around nucleotide 1400 and comprises around 5-10 or 10-20 contiguous nucleotides, especially a fragment greater than 17 nucleotides. In another embodiment the fragment is derived from around nucleotide 1400 to around nucleotide 1850 and comprises a continuous sequence of around 5-10, 10-20, or 20-25 nucleotides, especially a fragment greater than 23 nucleotides. In another embodiment, a fragment is derived from about nucleotide 1933 to about nucleotide 2421. Such a fragment comprises around 5-10 or 10-20 contiguous nucleotides. For the 26212 sulfatase, in one embodiment, a fragment is derived from around nucleotide 272 to around nucleotide 538 and comprises a contiguous sequence of around 5-10 or 10-20 nucleotides, especially a fragment greater than 17 nucleotides. In another embodiment, the fragment is derived from around nucleotide 538 to around nucleotide 751 and comprises a contiguous sequence of at least 5-10 or 10-20 nucleotides, especially greater than 12 nucleotides. In another embodiment, the fragment is derived from around nucleotide 1074 to around 1551 and comprises a contiguous nucleotide sequence of around 5-10, 10-20, or 20-30, especially greater than 20 nucleotides. In a further embodiment, the fragment is derived from around nucleotide 2052 to 2251 and comprises a contiguous sequence of 5-10 and 10-20 nucleotides, especially fragments greater than 18 nucleotides.

The fragment can comprise DNA or RNA and can be derived from either the coding or the non-coding sequence.

In another embodiment an isolated sulfatase nucleic acid encodes the entire coding region. In another embodiment the isolated sulfatase nucleic acid encodes a sequence corresponding to the mature protein. Other fragments include nucleotide sequences encoding the amino acid fragments described herein.

Thus, sulfatase nucleic acid fragments further include sequences corresponding to the regions described herein, subregions also described, and specific functional sites. Sulfatase nucleic acid fragments also include combinations of the regions, segments, motifs, and other functional sites described above. It is understood that a sulfatase fragment includes any nucleic acid sequence that does not include the entire gene. A person of ordinary skill in the art would be aware of the many permutations that are possible. Nucleic acid fragments, according to the present invention, are not to be construed as encompassing those fragments that may have been disclosed prior to the invention.

Where the location of the regions or sites have been predicted by computer analysis, one of ordinary skill would appreciate that the amino acid residues constituting these regions can vary depending on the criteria used to define the regions.

Polynucleotide Uses

The nucleotide sequences of the present invention can be used as a "query sequence" to perform a search against public databases, for example, to identify other family members or related sequences. For more information about public databases, see page 26, above.

The nucleic acid fragments of the invention provide probes or primers in assays such as those described below. "Probes" are oligonucleotides that hybridize in a base-specific manner to a complementary strand of nucleic acid. Such probes include polypeptide nucleic acids, as described in Nielsen *et al.* (1991) *Science 254*:1497-1500. Typically, a probe comprises a region of nucleotide sequence that hybridizes under highly stringent conditions to at least about 15, typically about 20-25, and more typically about 30, 40, 50 or 75 consecutive nucleotides of the nucleic acid sequence shown in SEQ ID NOS:2, 4, 6, 8, 11, 12, 13, or 14, and the

complements thereof. More typically, the probe further comprises a label, e.g., radioisotope, fluorescent compound, enzyme, or enzyme co-factor.

As used herein, the term "primer" refers to a single-stranded oligonucleotide which acts as a point of initiation of template-directed DNA synthesis using well-known methods (e.g., PCR, LCR) including, but not limited to those described herein. The appropriate length of the primer depends on the particular use, but typically ranges from about 15 to 30 nucleotides. The term "primer site" refers to the area of the target DNA to which a primer hybridizes. The term "primer pair" refers to a set of primers including a 5' (upstream) primer that hybridizes with the 5' end of the nucleic acid sequence to be amplified and a 3' (downstream) primer that hybridizes with the complement of the sequence to be amplified.

Sulfatase polynucleotides are thus useful for probes, primers, and in biological assays. Where the polynucleotides are used to assess sulfatase properties or functions, such as in the assays described herein, all or less than all of the entire cDNA can be useful. Assays specifically directed to sulfatase functions, such as assessing agonist or antagonist activity, encompass the use of known fragments. Further, diagnostic methods for assessing sulfatase function can also be practiced with any fragment, including those fragments that may have been known prior to the invention. Similarly, in methods involving treatment of sulfatase dysfunction, all fragments are encompassed including those, which may have been known in the art.

Sulfatase polynucleotides are useful as a hybridization probe for cDNA and genomic DNA to isolate a full-length cDNA and genomic clones encoding the polypeptides described in SEQ ID NOS:1, 3, 5, or 7, and to isolate cDNA and genomic clones that correspond to variants producing the same polypeptides shown in SEQ ID NOS:1, 3, 5, or 7, or the other variants described herein. Variants can be isolated from the same tissue and organism from which a polypeptide shown in SEQ ID NOS:1, 3, 5, or 7 was isolated, different tissues from the same organism, or from different organisms. This method is useful for isolating genes and cDNA that are developmentally-controlled and therefore may be expressed in the same tissue or different tissues at different points in the development of an organism.

The probe can correspond to any sequence along the entire length of the gene encoding the sulfatase polypeptide. Accordingly, it could be derived from 5' noncoding regions, the coding region, and 3' noncoding regions.

The nucleic acid probe can be, for example, the full-length cDNA of SEQ ID NOS:2, 4, 6, 8, 11, 12, 13, or 14 or a fragment thereof, such as an oligonucleotide of at least 5, 10, 15, 20, 25, 30, 50, 100, 250 or 500 nucleotides in length and sufficient to specifically hybridize under stringent conditions to mRNA or DNA.

Fragments of the polynucleotides described herein are also useful to synthesize larger fragments or full-length polynucleotides described herein, ribozymes or antisense molecules. For example, a fragment can be hybridized to any portion of an mRNA and a larger or full-length cDNA can be produced.

Antisense nucleic acids of the invention can be designed using the nucleotide sequences of SEQ ID NOS:2, 4, 6, 8, 11, 12, 13, or 14 and constructed using chemical synthesis and enzymatic ligation reactions using procedures known in the art. For example, an antisense nucleic acid (e.g., an antisense oligonucleotide) can be chemically synthesized using naturally occurring nucleotides or variously modified nucleotides designed to increase the biological stability of the molecules or to increase the physical stability of the duplex formed between the antisense and sense nucleic acids, e.g., phosphorothioate derivatives and acridine substituted nucleotides can be used. Examples of modified nucleotides which can be used to generate the antisense nucleic acid include 5-fluorouracil, 5-bromouracil, 5-chlorouracil, 5-iodouracil, hypoxanthine, xanthine, 4-acetylcytosine, 5-(carboxyhydroxylmethyl) uracil, 5-carboxymethylaminomethyl-2-thiouridine, 5-carboxymethylaminomethyluracil, dihydrouracil, beta-D-galactosylqueosine, inosine, N6-isopentenyladenine, 1-methylguanine, 1-methylinosine, 2,2-dimethylguanine, 2-methyladenine, 2-methylguanine, 3-methylcytosine, 5-methylcytosine, N6-adenine, 7-methylguanine, 5-methylaminomethyluracil, 5-methoxyaminomethyl-2-thiouracil, beta-D-mannosylqueosine, 5'-methoxycarboxymethyluracil, 5-methoxyuracil, 2-methylthio-N6-isopentenyladenine, uracil-5-oxyacetic acid (v), wybutoxosine, pseudouracil, queosine, 2-thiocytosine, 5-methyl-2-thiouracil, 2-thiouracil, 4-thiouracil, 5-methyluracil, uracil-5-oxyacetic acid methylester, uracil-5-oxyacetic acid (v), 5-methyl-2-thiouracil, 3-(3-amino-3-N-2-carboxypropyl) uracil, (acp3)w, and 2,6-

diaminopurine. Alternatively, the antisense nucleic acid can be produced biologically using an expression vector into which a nucleic acid has been subcloned in an antisense orientation (i.e., RNA transcribed from the inserted nucleic acid will be of an antisense orientation to a target nucleic acid of interest).

Additionally, the nucleic acid molecules of the invention can be modified at the base moiety, sugar moiety or phosphate backbone to improve, e.g., the stability, hybridization, or solubility of the molecule. For example, the deoxyribose phosphate backbone of the nucleic acids can be modified to generate peptide nucleic acids (see Hyrup *et al.* (1996) *Bioorganic & Medicinal Chemistry 4*:5). As used herein, the terms "peptide nucleic acids" or "PNAs" refer to nucleic acid mimics, e.g., DNA mimics, in which the deoxyribose phosphate backbone is replaced by a pseudopeptide backbone and only the four natural nucleobases are retained. The neutral backbone of PNAs has been shown to allow for specific hybridization to DNA and RNA under conditions of low ionic strength. The synthesis of PNA oligomers can be performed using standard solid phase peptide synthesis protocols as described in Hyrup *et al.* (1996), *supra*; Perry-O'Keefe *et al.* (1996) *Proc. Natl. Acad. Sci. USA 93*:14670. PNAs can be further modified, e.g., to enhance their stability, specificity or cellular uptake, by attaching lipophilic or other helper groups to PNA, by the formation of PNA-DNA chimeras, or by the use of liposomes or other techniques of drug delivery known in the art. The synthesis of PNA-DNA chimeras can be performed as described in Hyrup (1996), *supra*, Finn *et al.* (1996) *Nucleic Acids Res. 24*(17):3357-63, Mag *et al.* (1989) *Nucleic Acids Res. 17*:5973, and Peterser *et al.* (1975) *Bioorganic Med. Chem. Lett. 5*:1119.

The nucleic acid molecules and fragments of the invention can also include other appended groups such as peptides (e.g., for targeting host cell sulfatases *in vivo*), or agents facilitating transport across the cell membrane (see, e.g., Letsinger *et al.* (1989) *Proc. Natl. Acad. Sci. USA 86*:6553-6556; Lemaitre *et al.* (1987) *Proc. Natl. Acad. Sci. USA 84*:648-652; PCT Publication No. WO 88/0918) or the blood brain barrier (see, e.g., PCT Publication No. WO 89/10134). In addition, oligonucleotides can be modified with hybridization-triggered cleavage agents (see, e.g., Krol *et al.* (1988) *Bio-Techniques 6*:958-976) or intercalating agents (see, e.g., Zon (1988) *Pharm Res. 5*:539-549).

Sulfatase polynucleotides are also useful as primers for PCR to amplify any given region of a sulfatase polynucleotide.

Sulfatase polynucleotides are also useful for constructing recombinant vectors. Such vectors include expression vectors that express a portion of, or all of, the sulfatase polypeptides. Vectors also include insertion vectors, used to integrate into another polynucleotide sequence, such as into the cellular genome, to alter *in situ* expression of sulfatase genes and gene products. For example, an endogenous sulfatase coding sequence can be replaced via homologous recombination with all or part of the coding region containing one or more specifically introduced mutations.

Sulfatase polynucleotides are also useful for expressing antigenic portions of sulfatase proteins.

Sulfatase polynucleotides are also useful as probes for determining the chromosomal positions of sulfatase polynucleotides by means of *in situ* hybridization methods, such as FISH. (For a review of this technique, see Verma *et al.* (1988) *Human Chromosomes: A Manual of Basic Techniques* (Pergamon Press, New York), and PCR mapping of somatic cell hybrids. The mapping of the sequences to chromosomes is an important first step in correlating these sequences with genes associated with disease.

Reagents for chromosome mapping can be used individually to mark a single chromosome or a single site on that chromosome, or panels of reagents can be used for marking multiple sites and/or multiple chromosomes. Reagents corresponding to noncoding regions of the genes actually are preferred for mapping purposes. Coding sequences are more likely to be conserved within gene families, thus increasing the chance of cross hybridizations during chromosomal mapping.

Once a sequence has been mapped to a precise chromosomal location, the physical position of the sequence on the chromosome can be correlated with genetic map data. (Such data are found, for example, in V. McKusick, *Mendelian Inheritance in Man*, available on-line through Johns Hopkins University Welch Medical Library). The relationship between a gene and a disease mapped to the same chromosomal region, can then be identified through linkage analysis (co-inheritance of physically adjacent genes), described in, for example, Egeland *et al.* ((1987) *Nature 325*:783-787).

Moreover, differences in the DNA sequences between individuals affected and unaffected with a disease associated with a specified gene, can be determined. If a

mutation is observed in some or all of the affected individuals but not in any unaffected individuals, then the mutation is likely to be the causative agent of the particular disease. Comparison of affected and unaffected individuals generally involves first looking for structural alterations in the chromosomes, such as deletions or translocations, that are visible from chromosome spreads, or detectable using PCR based on that DNA sequence. Ultimately, complete sequencing of genes from several individuals can be performed to confirm the presence of a mutation and to distinguish mutations from polymorphisms.

Sulfatase polynucleotide probes are also useful to determine patterns of the presence of the gene encoding sulfatases and their variants with respect to tissue distribution, for example, whether gene duplication has occurred and whether the duplication occurs in all or only a subset of tissues. The genes can be naturally occurring or can have been introduced into a cell, tissue, or organism exogenously.

Sulfatase polynucleotides are also useful for designing ribozymes corresponding to all, or a part, of the mRNA produced from genes encoding the polynucleotides described herein.

Sulfatase polynucleotides are also useful for constructing host cells expressing a part, or all, of a sulfatase polynucleotide or polypeptide.

Sulfatase polynucleotides are also useful for constructing transgenic animals expressing all, or a part, of a sulfatase polynucleotide or polypeptide.

Sulfatase polynucleotides are also useful for making vectors that express part, or all, of a sulfatase polypeptide.

Sulfatase polynucleotides are also useful as hybridization probes for determining the level of sulfatase nucleic acid expression. Accordingly, the probes can be used to detect the presence of, or to determine levels of, sulfatase nucleic acid in cells, tissues, and in organisms. The nucleic acid whose level is determined can be DNA or RNA. Accordingly, probes corresponding to the polypeptides described herein can be used to assess gene copy number in a given cell, tissue, or organism. This is particularly relevant in cases in which there has been an amplification of a sulfatase gene.

Alternatively, the probe can be used in an *in situ* hybridization context to assess the position of extra copies of a sulfatase gene, as on extrachromosomal elements or as

integrated into chromosomes in which the sulfatase gene is not normally found, for example, as a homogeneously staining region.

These uses are relevant for diagnosis of disorders involving an increase or decrease in sulfatase expression relative to normal, such as a proliferative disorder, a differentiative or developmental disorder, or a hematopoietic disorder. Disorders in which sulfatase expression is relevant include, but are not limited to, those disclosed herein above.

Disorders in which 22438 sulfatase expression is relevant include, but are not limited to, those involving the tissues as disclosed herein and those associated with pain.

Disorders in which 23553 sulfatase expression is relevant include, but are not limited to, breast and colon carcinoma.

Disorders in which 25278 sulfatase expression is relevant include, but are not limited to, colon carcinoma.

Disorders in which 26212 sulfatase expression is relevant include, but are not limited to, hemangioma and uterine adenocarcinoma.

Thus, the present invention provides a method for identifying a disease or disorder associated with aberrant expression or activity of a sulfatase nucleic acid, in which a test sample is obtained from a subject and nucleic acid (e.g., mRNA, genomic DNA) is detected, wherein the presence of the nucleic acid is diagnostic for a subject having or at risk of developing a disease or disorder associated with aberrant expression or activity of the nucleic acid.

One aspect of the invention relates to diagnostic assays for determining nucleic acid expression as well as activity in the context of a biological sample (e.g., blood, serum, cells, tissue) to determine whether an individual has a disease or disorder, or is at risk of developing a disease or disorder, associated with aberrant nucleic acid expression or activity. Such assays can be used for prognostic or predictive purpose to thereby prophylactically treat an individual prior to the onset of a disorder characterized by or associated with expression or activity of the nucleic acid molecules.

In vitro techniques for detection of mRNA include Northern hybridizations and *in situ* hybridizations. *In vitro* techniques for detecting DNA includes Southern hybridizations and *in situ* hybridization.

Probes can be used as a part of a diagnostic test kit for identifying cells or tissues that express a sulfatase, such as by measuring the level of a sulfatase-encoding nucleic acid in a sample of cells from a subject e.g., mRNA or genomic DNA, or determining if the sulfatase gene has been mutated.

Nucleic acid expression assays are useful for drug screening to identify compounds that modulate sulfatase nucleic acid expression (e.g., antisense, polypeptides, peptidomimetics, small molecules or other drugs). A cell is contacted with a candidate compound and the expression of mRNA determined. The level of expression of the mRNA in the presence of the candidate compound is compared to the level of expression of the mRNA in the absence of the candidate compound. The candidate compound can then be identified as a modulator of nucleic acid expression based on this comparison and be used, for example to treat a disorder characterized by aberrant nucleic acid expression. The modulator can bind to the nucleic acid or indirectly modulate expression, such as by interacting with other cellular components that affect nucleic acid expression.

Modulatory methods can be performed *in vitro* (e.g., by culturing the cell with the agent) or, alternatively, *in vivo* (e.g., by administering the gent to a subject) in patients or in transgenic animals. The invention thus provides a method for identifying a compound that can be used to treat a disorder associated with nucleic acid expression of a sulfatase gene. The method typically includes assaying the ability of the compound to modulate the expression of the sulfatase nucleic acid and thus identifying a compound that can be used to treat a disorder characterized by undesired sulfatase nucleic acid expression.

The assays can be performed in cell-based and cell-free systems. Cell-based assays include cells naturally expressing the sulfatase nucleic acid or recombinant cells genetically engineered to express specific nucleic acid sequences. Alternatively, candidate compounds can be assayed *in vivo* in patients or in transgenic animals.

The assay for sulfatase nucleic acid expression can involve direct assay of nucleic acid levels, such as mRNA levels, or on collateral compounds (such as substrate

hydrolysis). Further, the expression of genes that are up- or down-regulated in response to sulfatase activity can also be assayed. In this embodiment the regulatory regions of these genes can be operably linked to a reporter gene such as luciferase.

Thus, modulators of sulfatase gene expression can be identified in a method wherein a cell is contacted with a candidate compound and the expression of mRNA determined. The level of expression of sulfatase mRNA in the presence of the candidate compound is compared to the level of expression of sulfatase mRNA in the absence of the candidate compound. The candidate compound can then be identified as a modulator of nucleic acid expression based on this comparison and be used, for example to treat a disorder characterized by aberrant nucleic acid expression. When expression of mRNA is statistically significantly greater in the presence of the candidate compound than in its absence, the candidate compound is identified as a stimulator of nucleic acid expression. When nucleic acid expression is statistically significantly less in the presence of the candidate compound than in its absence, the candidate compound is identified as an inhibitor of nucleic acid expression.

Accordingly, the invention provides methods of treatment, with the nucleic acid as a target, using a compound identified through drug screening as a gene modulator to modulate sulfatase nucleic acid expression. Modulation includes both up-regulation (i.e. activation or agonization) or down-regulation (suppression or antagonization) or effects on nucleic acid activity (e.g. when nucleic acid is mutated or improperly modified). Treatment is of disorders characterized by aberrant expression or activity of the nucleic acid.

Alternatively, a modulator for sulfatase nucleic acid expression can be a small molecule or drug identified using the screening assays described herein as long as the drug or small molecule inhibits sulfatase nucleic acid expression.

Sulfatase polynucleotides are also useful for monitoring the effectiveness of modulating compounds on the expression or activity of a sulfatase gene in clinical trials or in a treatment regimen. Thus, the gene expression pattern can serve as a barometer for the continuing effectiveness of treatment with the compound, particularly with compounds to which a patient can develop resistance. The gene expression pattern can also serve as a marker indicative of a physiological response of the affected cells to the compound. Accordingly, such monitoring would allow either increased administration

of the compound or the administration of alternative compounds to which the patient has not become resistant. Similarly, if the level of nucleic acid expression falls below a desirable level, administration of the compound could be commensurately decreased.

Monitoring can be, for example, as follows: (i) obtaining a pre-administration sample from a subject prior to administration of the agent; (ii) detecting the level of expression of a specified mRNA or genomic DNA of the invention in the pre-administration sample; (iii) obtaining one or more post-administration samples from the subject; (iv) detecting the level of expression or activity of the mRNA or genomic DNA in the post-administration samples; (v) comparing the level of expression or activity of the mRNA or genomic DNA in the pre-administration sample with the mRNA or genomic DNA in the post-administration sample or samples; and (vi) increasing or decreasing the administration of the agent to the subject accordingly.

Sulfatase polynucleotides are also useful in diagnostic assays for qualitative changes in sulfatase nucleic acid, and particularly in qualitative changes that lead to pathology. The polynucleotides can be used to detect mutations in sulfatase genes and gene expression products such as mRNA. The polynucleotides can be used as hybridization probes to detect naturally-occurring genetic mutations in a sulfatase gene and thereby to determine whether a subject with the mutation is at risk for a disorder caused by the mutation. Mutations include deletion, addition, or substitution of one or more nucleotides in the gene, chromosomal rearrangement, such as inversion or transposition, modification of genomic DNA, such as aberrant methylation patterns or changes in gene copy number, such as amplification. Detection of a mutated form of a sulfatase gene associated with a dysfunction provides a diagnostic tool for an active disease or susceptibility to disease when the disease results from overexpression, underexpression, or altered expression of a sulfatase.

Mutations in a sulfatase gene can be detected at the nucleic acid level by a variety of techniques. Genomic DNA can be analyzed directly or can be amplified by using PCR prior to analysis. RNA or cDNA can be used in the same way.

In certain embodiments, detection of the mutation involves the use of a probe/primer in a polymerase chain reaction (PCR) (see, e.g. U.S. Patent Nos. 4,683,195 and 4,683,202), such as anchor PCR or RACE PCR, or, alternatively, in a ligation chain reaction (LCR) (see, e.g., Landegran *et al.* (1988) *Science 241*:1077-1080; and

Nakazawa *et al.* (1994) *PNAS 91*:360-364), the latter of which can be particularly useful for detecting point mutations in the gene (see Abravaya *et al.* (1995) *Nucleic Acids Res. 23*:675-682). This method can include the steps of collecting a sample of cells from a patient, isolating nucleic acid (e.g., genomic, mRNA or both) from the cells of the sample, contacting the nucleic acid sample with one or more primers which specifically hybridize to a gene under conditions such that hybridization and amplification of the gene (if present) occurs, and detecting the presence or absence of an amplification product, or detecting the size of the amplification product and comparing the length to a control sample. Deletions and insertions can be detected by a change in size of the amplified product compared to the normal genotype. Point mutations can be identified by hybridizing amplified DNA to normal RNA or antisense DNA sequences.

It is anticipated that PCR and/or LCR may be desirable to use as a preliminary amplification step in conjunction with any of the techniques used for detecting mutations described herein.

Alternative amplification methods include: self sustained sequence replication (Guatelli *et al.* (1990) *Proc. Natl. Acad. Sci. USA 87*:1874-1878), transcriptional amplification system (Kwoh *et al.* (1989) *Proc. Natl. Acad. Sci. USA 86*:1173-1177), Q-Beta Replicase (Lizardi *et al.* (1988) *Bio/Technology 6*:1197), or any other nucleic acid amplification method, followed by the detection of the amplified molecules using techniques well-known to those of skill in the art. These detection schemes are especially useful for the detection of nucleic acid molecules if such molecules are present in very low numbers.

Alternatively, mutations in a sulfatase gene can be directly identified, for example, by alterations in restriction enzyme digestion patterns determined by gel electrophoresis.

Further, sequence-specific ribozymes (U.S. Patent No. 5,498,531) can be used to score for the presence of specific mutations by development or loss of a ribozyme cleavage site.

Perfectly matched sequences can be distinguished from mismatched sequences by nuclease cleavage digestion assays or by differences in melting temperature.

Sequence changes at specific locations can also be assessed by nuclease protection assays such as RNase and S1 protection or the chemical cleavage method.

Furthermore, sequence differences between a mutant sulfatase gene and a wild-type gene can be determined by direct DNA sequencing. A variety of automated sequencing procedures can be utilized when performing the diagnostic assays ((1995) *Biotechniques 19*:448), including sequencing by mass spectrometry (see, e.g., PCT International Publication No. WO 94/16101; Cohen *et al.* (1996) *Adv. Chromatogr. 36*:127-162; and Griffin *et al.* (1993) *Appl. Biochem. Biotechnol. 38*:147-159).

Other methods for detecting mutations in the gene include methods in which protection from cleavage agents is used to detect mismatched bases in RNA/RNA or RNA/DNA duplexes (Myers *et al.* (1985) *Science 230*:1242); Cotton *et al.* (1988) *PNAS 85*:4397; Saleeba *et al.* (1992) *Meth. Enzymol. 217*:286-295), electrophoretic mobility of mutant and wild type nucleic acid is compared (Orita *et al.* (1989) *PNAS 86*:2766; Cotton *et al. (1993) Mutat. Res. 285*:125-144; and Hayashi *et al.* (1992) *Genet. Anal. Tech. Appl. 9*:73-79), and movement of mutant or wild-type fragments in polyacrylamide gels containing a gradient of denaturant is assayed using denaturing gradient gel electrophoresis (Myers *et al.* (1985) *Nature 313*:495). The sensitivity of the assay may be enhanced by using RNA (rather than DNA), in which the secondary structure is more sensitive to a change in sequence. In one embodiment, the subject method utilizes heteroduplex analysis to separate double stranded heteroduplex molecules on the basis of changes in electrophoretic mobility (Keen *et al.* (1991) *Trends Genet. 7*:5). Examples of other techniques for detecting point mutations include, selective oligonucleotide hybridization, selective amplification, and selective primer extension.

In other embodiments, genetic mutations can be identified by hybridizing a sample and control nucleic acids, e.g., DNA or RNA, to high density arrays containing hundreds or thousands of oligonucleotide probes (Cronin *et al.* (1996) *Human Mutation 7*:244-255; Kozal *et al.* (1996) *Nature Medicine 2*:753-759). For example, genetic mutations can be identified in two dimensional arrays containing light-generated DNA probes as described in Cronin *et al. supra.* Briefly, a first hybridization array of probes can be used to scan through long stretches of DNA in a sample and control to identify base changes between the sequences by making linear arrays of sequential overlapping probes. This step allows the identification of point mutations. This step is followed by a second hybridization array that allows the

characterization of specific mutations by using smaller, specialized probe arrays complementary to all variants or mutations detected. Each mutation array is composed of parallel probe sets, one complementary to the wild-type gene and the other complementary to the mutant gene.

Sulfatase polynucleotides are also useful for testing an individual for a genotype that while not necessarily causing the disease, nevertheless affects the treatment modality. Thus, the polynucleotides can be used to study the relationship between an individual's genotype and the individual's response to a compound used for treatment (pharmacogenomic relationship). In the present case, for example, a mutation in the sulfatase gene that results in altered affinity for a substrate-related compound could result in an excessive or decreased drug effect with standard concentrations of the compound. Accordingly, the sulfatase polynucleotides described herein can be used to assess the mutation content of the gene in an individual in order to select an appropriate compound or dosage regimen for treatment.

Thus polynucleotides displaying genetic variations that affect treatment provide a diagnostic target that can be used to tailor treatment in an individual. Accordingly, the production of recombinant cells and animals containing these polymorphisms allow effective clinical design of treatment compounds and dosage regimens.

The methods can involve obtaining a control biological sample from a control subject, contacting the control sample with a compound or agent capable of detecting mRNA, or genomic DNA, such that the presence of mRNA or genomic DNA is detected in the biological sample, and comparing the presence of mRNA or genomic DNA in the control sample with the presence of mRNA or genomic DNA in the test sample.

Sulfatase polynucleotides are also useful for chromosome identification when the sequence is identified with an individual chromosome and to a particular location on the chromosome. First, the DNA sequence is matched to the chromosome by *in situ* or other chromosome-specific hybridization. Sequences can also be correlated to specific chromosomes by preparing PCR primers that can be used for PCR screening of somatic cell hybrids containing individual chromosomes from the desired species. Only hybrids containing the chromosome containing the gene homologous to the primer will yield an amplified fragment. Sublocalization can be achieved using chromosomal fragments.

Other strategies include prescreening with labeled flow-sorted chromosomes and preselection by hybridization to chromosome-specific libraries. Further mapping strategies include fluorescence *in situ* hybridization, which allows hybridization with probes shorter than those traditionally used. Reagents for chromosome mapping can be used individually to mark a single chromosome or a single site on the chromosome, or panels of reagents can be used for marking multiple sites and/or multiple chromosomes. Reagents corresponding to noncoding regions of the genes actually are preferred for mapping purposes. Coding sequences are more likely to be conserved within gene families, thus increasing the chance of cross hybridizations during chromosomal mapping.

Sulfatase polynucleotides can also be used to identify individuals from small biological samples. This can be done for example using restriction fragment-length polymorphism (RFLP) to identify an individual. Thus, the polynucleotides described herein are useful as DNA markers for RFLP (See U.S. Patent No. 5,272,057).

Furthermore, the sulfatase sequences can be used to provide an alternative technique, which determines the actual DNA sequence of selected fragments in the genome of an individual. Thus, the sulfatase sequences described herein can be used to prepare two PCR primers from the 5' and 3' ends of the sequences. These primers can then be used to amplify DNA from an individual for subsequent sequencing.

Panels of corresponding DNA sequences from individuals prepared in this manner can provide unique individual identifications, as each individual will have a unique set of such DNA sequences. It is estimated that allelic variation in humans occurs with a frequency of about once per each 500 bases. Allelic variation occurs to some degree in the coding regions of these sequences, and to a greater degree in the noncoding regions. Sulfatase sequences can be used to obtain such identification sequences from individuals and from tissue. The sequences represent unique fragments of the human genome. Each of the sequences described herein can, to some degree, be used as a standard against which DNA from an individual can be compared for identification purposes.

If a panel of reagents from the sequences is used to generate a unique identification database for an individual, those same reagents can later be used to identify tissue from that individual. Using the unique identification database, positive

identification of the individual, living or dead, can be made from extremely small tissue samples.

Sulfatase polynucleotides can also be used in forensic identification procedures. PCR technology can be used to amplify DNA sequences taken from very small biological samples, such as a single hair follicle, body fluids (e.g. blood, saliva, or semen). The amplified sequence can then be compared to a standard allowing identification of the origin of the sample.

Sulfatase polynucleotides can thus be used to provide polynucleotide reagents, e.g., PCR primers, targeted to specific loci in the human genome, which can enhance the reliability of DNA-based forensic identifications by, for example, providing another "identification marker" (i.e. another DNA sequence that is unique to a particular individual). As described above, actual base sequence information can be used for identification as an accurate alternative to patterns formed by restriction enzyme generated fragments. Sequences targeted to the noncoding region are particularly useful since greater polymorphism occurs in the noncoding regions, making it easier to differentiate individuals using this technique.

Sulfatase polynucleotides can further be used to provide polynucleotide reagents, e.g., labeled or labelable probes which can be used in, for example, an *in situ* hybridization technique, to identify a specific tissue. This is useful in cases in which a forensic pathologist is presented with a tissue of unknown origin. Panels of sulfatase probes can be used to identify tissue by species and/or by organ type.

In a similar fashion, these primers and probes can be used to screen tissue culture for contamination (i.e. screen for the presence of a mixture of different types of cells in a culture).

Alternatively, sulfatase polynucleotides can be used directly to block transcription or translation of sulfatase gene sequences by means of antisense or ribozyme constructs. Thus, in a disorder characterized by abnormally high or undesirable sulfatase gene expression, nucleic acids can be directly used for treatment.

Sulfatase polynucleotides are thus useful as antisense constructs to control sulfatase gene expression in cells, tissues, and organisms. A DNA antisense polynucleotide is designed to be complementary to a region of the gene involved in transcription, preventing transcription and hence production of sulfatase protein. An

antisense RNA or DNA polynucleotide would hybridize to the mRNA and thus block translation of mRNA into sulfatase protein.

Examples of antisense molecules useful to inhibit nucleic acid expression include antisense molecules complementary to a fragment of the 5' untranslated region of SEQ ID NOS:2, 4, 6, or 8, which also includes the start codon and antisense molecules which are complementary to a fragment of the 3' untranslated region of SEQ ID NOS:2, 4, 6, or 8.

Alternatively, a class of antisense molecules can be used to inactivate mRNA in order to decrease expression of sulfatase nucleic acid. Accordingly, these molecules can treat a disorder characterized by abnormal or undesired sulfatase nucleic acid expression. This technique involves cleavage by means of ribozymes containing nucleotide sequences complementary to one or more regions in the mRNA that attenuate the ability of the mRNA to be translated. Possible regions include coding regions and particularly coding regions corresponding to the catalytic and other functional activities of the sulfatase protein.

Sulfatase polynucleotides also provide vectors for gene therapy in patients containing cells that are aberrant in sulfatase gene expression. Thus, recombinant cells, which include the patient's cells that have been engineered *ex vivo* and returned to the patient, are introduced into an individual where the cells produce the desired sulfatase protein to treat the individual.

The invention also encompasses kits for detecting the presence of a sulfatase nucleic acid in a biological sample. For example, the kit can comprise reagents such as a labeled or labelable nucleic acid or agent capable of detecting sulfatase nucleic acid in a biological sample; means for determining the amount of sulfatase nucleic acid in the sample; and means for comparing the amount of sulfatase nucleic acid in the sample with a standard. The compound or agent can be packaged in a suitable container. The kit can further comprise instructions for using the kit to detect sulfatase mRNA or DNA.

Computer Readable Means

The nucleotide or amino acid sequences of the invention are also provided in a variety of mediums to facilitate use thereof. As used herein, "provided" refers to a manufacture, other than an isolated nucleic acid or amino acid molecule, which

contains a nucleotide or amino acid sequence of the present invention. Such a manufacture provides the nucleotide or amino acid sequences, or a subset thereof (e.g., a subset of open reading frames (ORFs)) in a form which allows a skilled artisan to examine the manufacture using means not directly applicable to examining the nucleotide or amino acid sequences, or a subset thereof, as they exists in nature or in purified form.

In one application of this embodiment, a nucleotide or amino acid sequence of the present invention can be recorded on computer readable media. As used herein, "computer readable media" refers to any medium that can be read and accessed directly by a computer. Such media include, but are not limited to: magnetic storage media, such as floppy discs, hard disc storage medium, and magnetic tape; optical storage media such as CD-ROM; electrical storage media such as RAM and ROM; and hybrids of these categories such as magnetic/optical storage media. The skilled artisan will readily appreciate how any of the presently known computer readable mediums can be used to create a manufacture comprising computer readable medium having recorded thereon a nucleotide or amino acid sequence of the present invention.

As used herein, "recorded" refers to a process for storing information on computer readable medium. The skilled artisan can readily adopt any of the presently known methods for recording information on computer readable medium to generate manufactures comprising the nucleotide or amino acid sequence information of the present invention.

A variety of data storage structures are available to a skilled artisan for creating a computer readable medium having recorded thereon a nucleotide or amino acid sequence of the present invention. The choice of the data storage structure will generally be based on the means chosen to access the stored information. In addition, a variety of data processor programs and formats can be used to store the nucleotide sequence information of the present invention on computer readable medium. The sequence information can be represented in a word processing text file, formatted in commercially-available software such as WordPerfect and Microsoft Word, or represented in the form of an ASCII file, stored in a database application, such as DB2, Sybase, Oracle, or the like. The skilled artisan can readily adapt any number of dataprocessor structuring formats (e.g., text file or database) in order to obtain

computer readable medium having recorded thereon the nucleotide sequence information of the present invention.

By providing the nucleotide or amino acid sequences of the invention in computer readable form, the skilled artisan can routinely access the sequence information for a variety of purposes. For example, one skilled in the art can use the nucleotide or amino acid sequences of the invention in computer readable form to compare a target sequence or target structural motif with the sequence information stored within the data storage means. Search means are used to identify fragments or regions of the sequences of the invention which match a particular target sequence or target motif.

As used herein, a "target sequence" can be any DNA or amino acid sequence of six or more nucleotides or two or more amino acids. A skilled artisan can readily recognize that the longer a target sequence is, the less likely a target sequence will be present as a random occurrence in the database. The most preferred sequence length of a target sequence is from about 10 to 100 amino acids or from about 30 to 300 nucleotide residues. However, it is well recognized that commercially important fragments, such as sequence fragments involved in gene expression and protein processing, may be of shorter length.

As used herein, "a target structural motif," or "target motif," refers to any rationally selected sequence or combination of sequences in which the sequence(s) are chosen based on a three-dimensional configuration which is formed upon the folding of the target motif. There are a variety of target motifs known in the art. Protein target motifs include, but are not limited to, enzyme active sites and signal sequences. Nucleic acid target motifs include, but are not limited to, promoter sequences, hairpin structures and inducible expression elements (protein binding sequences).

Computer software is publicly available which allows a skilled artisan to access sequence information provided in a computer readable medium for analysis and comparison to other sequences. A variety of known algorithms are disclosed publicly and a variety of commercially available software for conducting search means are and can be used in the computer-based systems of the present invention. Examples of such software includes, but is not limited to, MacPattern (EMBL), BLASTN and BLASTX (NCBIA).

For example, software which implements the BLAST (Altschul *et al.* (1990) *J. Mol. Biol. 215*:403-410) and BLAZE (Brutlag *et al.* (1993) *Comp. Chem. 17*:203-207) search algorithms on a Sybase system can be used to identify open reading frames (ORFs) of the sequences of the invention which contain homology to ORFs or proteins from other libraries. Such ORFs are protein encoding fragments and are useful in producing commercially important proteins such as enzymes used in various reactions and in the production of commercially useful metabolites.

Vectors/Host Cells

The invention also provides vectors containing sulfatase polynucleotides. The term "vector" refers to a vehicle, preferably a nucleic acid molecule that can transport sulfatase polynucleotides. When the vector is a nucleic acid molecule, the sulfatase polynucleotides are covalently linked to the vector nucleic acid. With this aspect of the invention, the vector includes a plasmid, single or double stranded phage, a single or double stranded RNA or DNA viral vector, or artificial chromosome, such as a BAC, PAC, YAC, OR MAC.

A vector can be maintained in the host cell as an extrachromosomal element where it replicates and produces additional copies of sulfatase polynucleotides. Alternatively, the vector may integrate into the host cell genome and produce additional copies of sulfatase polynucleotides when the host cell replicates.

The invention provides vectors for the maintenance (cloning vectors) or vectors for expression (expression vectors) of sulfatase polynucleotides. The vectors can function in prokaryotic or eukaryotic cells or in both (shuttle vectors).

Expression vectors contain cis-acting regulatory regions that are operably linked in the vector to sulfatase polynucleotides such that transcription of the polynucleotides is allowed in a host cell. The polynucleotides can be introduced into the host cell with a separate polynucleotide capable of affecting transcription. Thus, the second polynucleotide may provide a trans-acting factor interacting with the cis-regulatory control region to allow transcription of sulfatase polynucleotides from the vector. Alternatively, a trans-acting factor may be supplied by the host cell. Finally, a trans-acting factor can be produced from the vector itself.

It is understood, however, that in some embodiments, transcription and/or translation of sulfatase polynucleotides can occur in a cell-free system.

The regulatory sequence to which the polynucleotides described herein can be operably linked include promoters for directing mRNA transcription. These include, but are not limited to, the left promoter from bacteriophage λ, the lac, TRP, and TAC promoters from *E. coli*, the early and late promoters from SV40, the CMV immediate early promoter, the adenovirus early and late promoters, and retrovirus long-terminal repeats.

In addition to control regions that promote transcription, expression vectors may also include regions that modulate transcription, such as repressor binding sites and enhancers. Examples include the SV40 enhancer, the cytomegalovirus immediate early enhancer, polyoma enhancer, adenovirus enhancers, and retrovirus LTR enhancers.

In addition to containing sites for transcription initiation and control, expression vectors can also contain sequences necessary for transcription termination and, in the transcribed region a ribosome binding site for translation. Other regulatory control elements for expression include initiation and termination codons as well as polyadenylation signals. The person of ordinary skill in the art would be aware of the numerous regulatory sequences that are useful in expression vectors. Such regulatory sequences are described, for example, in Sambrook *et al.* (1989) *Molecular Cloning: A Laboratory Manual 2nd. ed.*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY).

A variety of expression vectors can be used to express a sulfatase polynucleotide. Such vectors include chromosomal, episomal, and virus-derived vectors, for example vectors derived from bacterial plasmids, from bacteriophage, from yeast episomes, from yeast chromosomal elements, including yeast artificial chromosomes, from viruses such as baculoviruses, papovaviruses such as SV40, Vaccinia viruses, adenoviruses, poxviruses, pseudorabies viruses, and retroviruses. Vectors may also be derived from combinations of these sources such as those derived from plasmid and bacteriophage genetic elements, e.g. cosmids and phagemids. Appropriate cloning and expression vectors for prokaryotic and eukaryotic hosts are described in Sambrook *et al.* (1989) *Molecular Cloning: A Laboratory Manual 2nd. ed.*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

The regulatory sequence may provide constitutive expression in one or more host cells (i.e. tissue specific) or may provide for inducible expression in one or more cell types such as by temperature, nutrient additive, or exogenous factor such as a hormone or other ligand. A variety of vectors providing for constitutive and inducible expression in prokaryotic and eukaryotic hosts are well known to those of ordinary skill in the art.

Sulfatase polynucleotides can be inserted into the vector nucleic acid by well-known methodology. Generally, the DNA sequence that will ultimately be expressed is joined to an expression vector by cleaving the DNA sequence and the expression vector with one or more restriction enzymes and then ligating the fragments together. Procedures for restriction enzyme digestion and ligation are well known to those of ordinary skill in the art.

The vector containing the appropriate polynucleotide can be introduced into an appropriate host cell for propagation or expression using well-known techniques. Bacterial cells include, but are not limited to, *E. coli*, *Streptomyces*, and *Salmonella typhimurium*. Eukaryotic cells include, but are not limited to, yeast, insect cells such as *Drosophila*, animal cells such as COS and CHO cells, and plant cells.

As described herein, it may be desirable to express the polypeptide as a fusion protein. Accordingly, the invention provides fusion vectors that allow for the production of sulfatase polypeptides. Fusion vectors can increase the expression of a recombinant protein, increase the solubility of the recombinant protein, and aid in the purification of the protein by acting for example as a ligand for affinity purification. A proteolytic cleavage site may be introduced at the junction of the fusion moiety so that the desired polypeptide can ultimately be separated from the fusion moiety. Proteolytic enzymes include, but are not limited to, factor Xa, thrombin, and enterokinase. Typical fusion expression vectors include pGEX (Smith *et al.* (1988) *Gene 67*:31-40), pMAL (New England Biolabs, Beverly, MA) and pRIT5 (Pharmacia, Piscataway, NJ) which fuse glutathione S-transferase (GST), maltose E binding protein, or protein A, respectively, to the target recombinant protein. Examples of suitable inducible non-fusion *E. coli* expression vectors include pTrc (Amann *et al.* (1988) *Gene 69*:301-315) and pET 11d (Studier *et al.* (1990) *Gene Expression Technology: Methods in Enzymology 185*:60-89).

Recombinant protein expression can be maximized in a host bacteria by providing a genetic background wherein the host cell has an impaired capacity to

proteolytically cleave the recombinant protein. (Gottesman, S. (1990) *Gene Expression Technology: Methods in Enzymology* 185, Academic Press, San Diego, California 119-128).

It is further recognized that the nucleic acid sequences of the invention can be altered to contain codons, which are preferred, or non preferred, for a particular expression system. For example, the nucleic acid can be one in which at least one altered codon, and preferably at least 10%, or 20% of the codons have been altered such that the sequence is optimized for expression in *E. coli*, yeast, human, insect, or CHO cells. Methods for determining such codon usage are well known in the art.

Sulfatase polynucleotides can also be expressed by expression vectors that are operative in yeast. Examples of vectors for expression in yeast e.g., *S. cerevisiae* include pYepSec1 (Baldari *et al.* (1987) *EMBO J. 6*:229-234), pMFa (Kurjan *et al.* (1982) *Cell 30*:933-943), pJRY88 (Schultz *et al.* (1987) *Gene 54*:113-123), and pYES2 (Invitrogen Corporation, San Diego, CA).

Sulfatase polynucleotides can also be expressed in insect cells using, for example, baculovirus expression vectors. Baculovirus vectors available for expression of proteins in cultured insect cells (e.g., Sf 9 cells) include the pAc series (Smith *et al.* (1983) *Mol. Cell Biol. 3*:2156-2165) and the pVL series (Lucklow *et al.* (1989) *Virology 170*:31-39).

In certain embodiments of the invention, the polynucleotides described herein are expressed in mammalian cells using mammalian expression vectors. Examples of mammalian expression vectors include pCDM8 (Seed, B. (1987) *Nature 329*:840) and pMT2PC (Kaufman *et al.* (1987) *EMBO J. 6*:187-195).

The expression vectors listed herein are provided by way of example only of the well-known vectors available to those of ordinary skill in the art that would be useful to express sulfatase polynucleotides. The person of ordinary skill in the art would be aware of other vectors suitable for maintenance propagation or expression of the polynucleotides described herein. These are found for example in Sambrook *et al.* (1989) *Molecular Cloning: A Laboratory Manual 2nd, ed., Cold Spring Harbor Laboratory,* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

The invention also encompasses vectors in which the nucleic acid sequences described herein are cloned into the vector in reverse orientation, but operably linked to a

regulatory sequence that permits transcription of antisense RNA. Thus, an antisense transcript can be produced to all, or to a portion, of the polynucleotide sequences described herein, including both coding and non-coding regions. Expression of this antisense RNA is subject to each of the parameters described above in relation to expression of the sense RNA (regulatory sequences, constitutive or inducible expression, tissue-specific expression).

The invention also relates to recombinant host cells containing the vectors described herein. Host cells therefore include prokaryotic cells, lower eukaryotic cells such as yeast, other eukaryotic cells such as insect cells, and higher eukaryotic cells such as mammalian cells.

The recombinant host cells are prepared by introducing the vector constructs described herein into the cells by techniques readily available to the person of ordinary skill in the art. These include, but are not limited to, calcium phosphate transfection, DEAE-dextran-mediated transfection, cationic lipid-mediated transfection, electroporation, transduction, infection, lipofection, and other techniques such as those found in Sambrook *et al.* (*Molecular Cloning: A Laboratory Manual,* 2d ed., Cold Spring Harbor Laboratory, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY).

Host cells can contain more than one vector. Thus, different nucleotide sequences can be introduced on different vectors of the same cell. Similarly, sulfatase polynucleotides can be introduced either alone or with other polynucleotides that are not related to sulfatase polynucleotides such as those providing trans-acting factors for expression vectors. When more than one vector is introduced into a cell, the vectors can be introduced independently, co-introduced or joined to the sulfatase polynucleotide vector.

In the case of bacteriophage and viral vectors, these can be introduced into cells as packaged or encapsulated virus by standard procedures for infection and transduction. Viral vectors can be replication-competent or replication-defective. In the case in which viral replication is defective, replication will occur in host cells providing functions that complement the defects.

Vectors generally include selectable markers that enable the selection of the subpopulation of cells that contain the recombinant vector constructs. The marker can

be contained in the same vector that contains the polynucleotides described herein or may be on a separate vector. Markers include tetracycline or ampicillin-resistance genes for prokaryotic host cells and dihydrofolate reductase or neomycin resistance for eukaryotic host cells. However, any marker that provides selection for a phenotypic trait will be effective.

While the mature proteins can be produced in bacteria, yeast, mammalian cells, and other cells under the control of the appropriate regulatory sequences, cell-free transcription and translation systems can also be used to produce these proteins using RNA derived from the DNA constructs described herein.

Where secretion of the polypeptide is desired, appropriate secretion signals are incorporated into the vector. The signal sequence can be endogenous to the sulfatase polypeptides or heterologous to these polypeptides.

Where the polypeptide is not secreted into the medium, the protein can be isolated from the host cell by standard disruption procedures, including freeze thaw, sonication, mechanical disruption, use of lysing agents and the like. The polypeptide can then be recovered and purified by well-known purification methods including ammonium sulfate precipitation, acid extraction, anion or cationic exchange chromatography, phosphocellulose chromatography, hydrophobic-interaction chromatography, affinity chromatography, hydroxylapatite chromatography, lectin chromatography, or high performance liquid chromatography.

It is also understood that depending upon the host cell in recombinant production of the polypeptides described herein, the polypeptides can have various glycosylation patterns, depending upon the cell, or maybe non-glycosylated as when produced in bacteria. In addition, the polypeptides may include an initial modified methionine in some cases as a result of a host-mediated process.

Uses of Vectors and Host Cells

It is understood that "host cells" and "recombinant host cells" refer not only to the particular subject cell but also to the progeny or potential progeny of such a cell. Because certain modifications may occur in succeeding generations due to either mutation or environmental influences, such progeny may not, in fact, be identical to the parent cell, but are still included within the scope of the term as used herein. A

"purified preparation of cells", as used herein, refers to, in the case of plant or animal cells, an *in vitro* preparation of cells and not an entire intact plant or animal. In the case of cultured cells or microbial cells, it consists of a preparation of at least 10% and more preferably 50% of the subject cells.

The host cells expressing the polypeptides described herein, and particularly recombinant host cells, have a variety of uses. First, the cells are useful for producing sulfatase proteins or polypeptides that can be further purified to produce desired amounts of sulfatase protein or fragments. Thus, host cells containing expression vectors are useful for polypeptide production.

Host cells are also useful for conducting cell-based assays involving sulfatase or sulfatase fragments. Thus, a recombinant host cell expressing a native sulfatase is useful to assay for compounds that stimulate or inhibit sulfatase function, gene expression at the level of transcription or translation, and interaction with other cellular components.

Host cells are also useful for identifying sulfatase mutants in which these functions are affected. If the mutants naturally occur and give rise to a pathology, host cells containing the mutations are useful to assay compounds that have a desired effect on the mutant sulfatase (for example, stimulating or inhibiting function) which may not be indicated by their effect on the native sulfatase.

Recombinant host cells are also useful for expressing the chimeric polypeptides described herein to assess compounds that activate or suppress activation by means of a heterologous domain, segment, site, and the like, as disclosed herein.

Further, mutant sulfatases can be designed in which one or more of the various functions is engineered to be increased or decreased and used to augment or replace sulfatase proteins in an individual. Thus, host cells can provide a therapeutic benefit by replacing an aberrant sulfatase or providing an aberrant sulfatase that provides a therapeutic result. In one embodiment, the cells provide sulfatases that are abnormally active.

In another embodiment, the cells provide sulfatases that are abnormally inactive. These sulfatases can compete with endogenous sulfatases in the individual.

In another embodiment, cells expressing sulfatases that cannot be activated, are introduced into an individual in order to compete with endogenous sulfatases for substrate. For example, in the case in which excessive substrate or substrate analog is

part of a treatment modality, it may be necessary to effectively inactivate the substrate or substrate analog at a specific point in treatment. Providing cells that compete for the molecule, but which cannot be affected by sulfatase activation would be beneficial.

Homologously recombinant host cells can also be produced that allow the *in situ* alteration of endogenous sulfatase polynucleotide sequences in a host cell genome. The host cell includes, but is not limited to, a stable cell line, cell *in vivo*, or cloned microorganism. This technology is more fully described in WO 93/09222, WO 91/12650, WO 91/06667, U.S. 5,272,071, and U.S. 5,641,670. Briefly, specific polynucleotide sequences corresponding to the sulfatase polynucleotides or sequences proximal or distal to a sulfatase gene are allowed to integrate into a host cell genome by homologous recombination where expression of the gene can be affected. In one embodiment, regulatory sequences are introduced that either increase or decrease expression of an endogenous sequence. Accordingly, a sulfatase protein can be produced in a cell not normally producing it. Alternatively, increased expression of sulfatase protein can be effected in a cell normally producing the protein at a specific level. Further, expression can be decreased or eliminated by introducing a specific regulatory sequence. The regulatory sequence can be heterologous to the sulfatase protein sequence or can be a homologous sequence with a desired mutation that affects expression. Alternatively, the entire gene can be deleted. The regulatory sequence can be specific to the host cell or capable of functioning in more than one cell type. Still further, specific mutations can be introduced into any desired region of the gene to produce mutant sulfatase proteins. Such mutations could be introduced, for example, into the specific functional regions such as the peptide substrate-binding site.

In one embodiment, the host cell can be a fertilized oocyte or embryonic stem cell that can be used to produce a transgenic animal containing the altered sulfatase gene. Alternatively, the host cell can be a stem cell or other early tissue precursor that gives rise to a specific subset of cells and can be used to produce transgenic tissues in an animal. See also Thomas *et al.*, *Cell 51*:503 (1987) for a description of homologous recombination vectors. The vector is introduced into an embryonic stem cell line (e.g., by electroporation) and cells in which the introduced gene has homologously recombined with the endogenous sulfatase gene is selected (see e.g., Li, E. *et al.* (1992) *Cell 69*:915). The selected cells are then injected into a blastocyst of an animal (e.g., a

mouse) to form aggregation chimeras (see e.g., Bradley, A. in *Teratocarcinomas and Embryonic Stem Cells: A Practical Approach*, E.J. Robertson, ed. (IRL, Oxford, 1987) pp. 113-152). A chimeric embryo can then be implanted into a suitable pseudopregnant female foster animal and the embryo brought to term. Progeny harboring the homologously recombined DNA in their germ cells can be used to breed animals in which all cells of the animal contain the homologously recombined DNA by germline transmission of the transgene. Methods for constructing homologous recombination vectors and homologous recombinant animals are described further in Bradley, A. (1991) *Current Opinions in Biotechnology* 2:823-829 and in PCT International Publication Nos. WO 90/11354; WO 91/01140; and WO 93/04169.

The genetically engineered host cells can be used to produce non-human transgenic animals. A transgenic animal is preferably a mammal, for example a rodent, such as a rat or mouse, in which one or more of the cells of the animal include a transgene. A transgene is exogenous DNA which is integrated into the genome of a cell from which a transgenic animal develops and which remains in the genome of the mature animal in one or more cell types or tissues of the transgenic animal. These animals are useful for studying the function of a sulfatase protein and identifying and evaluating modulators of sulfatase protein activity.

Other examples of transgenic animals include non-human primates, sheep, dogs, cows, goats, chickens, and amphibians.

In one embodiment, a host cell is a fertilized oocyte or an embryonic stem cell into which sulfatase polynucleotide sequences have been introduced.

A transgenic animal can be produced by introducing nucleic acid into the male pronuclei of a fertilized oocyte, e.g., by microinjection, retroviral infection, and allowing the oocyte to develop in a pseudopregnant female foster animal. Any of the sulfatase nucleotide sequences can be introduced as a transgene into the genome of a non-human animal, such as a mouse.

Any of the regulatory or other sequences useful in expression vectors can form part of the transgenic sequence. This includes intronic sequences and polyadenylation signals, if not already included. A tissue-specific regulatory sequence(s) can be operably linked to the transgene to direct expression of the sulfatase protein to particular cells.

Methods for generating transgenic animals via embryo manipulation and microinjection, particularly animals such as mice, have become conventional in the art and are described, for example, in U.S. Patent Nos. 4,736,866 and 4,870,009, both by Leder *et al.*, U.S. Patent No. 4,873,191 by Wagner *et al.* and in Hogan, B., *Manipulating the Mouse Embryo*, (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y., 1986). Similar methods are used for production of other transgenic animals. A transgenic founder animal can be identified based upon the presence of the transgene in its genome and/or expression of transgenic mRNA in tissues or cells of the animals. A transgenic founder animal can then be used to breed additional animals carrying the transgene. Moreover, transgenic animals carrying a transgene can further be bred to other transgenic animals carrying other transgenes. A transgenic animal also includes animals in which the entire animal or tissues in the animal have been produced using the homologously recombinant host cells described herein.

In another embodiment, transgenic non-human animals can be produced which contain selected systems, which allow for regulated expression of the transgene. One example of such a system is the *cre/loxP* recombinase system of bacteriophage P1. For a description of the *cre/loxP* recombinase system, see, e.g., Lakso *et al.* (1992) *PNAS 89*:6232-6236. Another example of a recombinase system is the FLP recombinase system of *S. cerevisiae* (O'Gorman *et al.* (1991) *Science 251*:1351-1355. If a *cre/loxP* recombinase system is used to regulate expression of the transgene, animals containing transgenes encoding both the Cre recombinase and a selected protein is required. Such animals can be provided through the construction of "double" transgenic animals, e.g., by mating two transgenic animals, one containing a transgene encoding a selected protein and the other containing a transgene encoding a recombinase.

Clones of the non-human transgenic animals described herein can also be produced according to the methods described in Wilmut *et al.* (1997) *Nature 385*:810-813 and PCT International Publication Nos. WO 97/07668 and WO 97/07669. In brief, a cell, e.g., a somatic cell, from the transgenic animal can be isolated and induced to exit the growth cycle and enter G_o phase. The quiescent cell can then be fused, e.g., through the use of electrical pulses, to an enucleated oocyte from an animal of the same species from which the quiescent cell is isolated. The reconstructed oocyte is then cultured such that it develops to morula or blastocyst and then transferred to a pseudopregnant female

foster animal. The offspring born of this female animal will be a clone of the animal from which the cell, e.g., the somatic cell, is isolated.

Transgenic animals containing recombinant cells that express the polypeptides described herein are useful to conduct the assays described herein in an *in vivo* context. Accordingly, the various physiological factors that are present *in vivo* and that could affect binding or activation, may not be evident from *in vitro* cell-free or cell-based assays. Accordingly, it is useful to provide non-human transgenic animals to assay *in vivo* sulfatase function, including peptide interaction, the effect of specific mutant sulfatases on sulfatase function and peptide interaction, and the effect of chimeric sulfatases. It is also possible to assess the effect of null mutations, that is mutations that substantially or completely eliminate one or more sulfatase functions.

In general, methods for producing transgenic animals include introducing a nucleic acid sequence according to the present invention, the nucleic acid sequence capable of expressing the protein in a transgenic animal, into a cell in culture or *in vivo*. When introduced *in vivo*, the nucleic acid is introduced into an intact organism such that one or more cell types and, accordingly, one or more tissue types, express the nucleic acid encoding the protein. Alternatively, the nucleic acid can be introduced into virtually all cells in an organism by transfecting a cell in culture, such as an embryonic stem cell, as described herein for the production of transgenic animals, and this cell can be used to produce an entire transgenic organism. As described, in a further embodiment, the host cell can be a fertilized oocyte. Such cells are then allowed to develop in a female foster animal to produce the transgenic organism.

Pharmaceutical Compositions

Sulfatase nucleic acid molecules, proteins, modulators of the protein, and antibodies (also referred to herein as "active compounds") can be incorporated into pharmaceutical compositions suitable for administration to a subject, e.g., a human. Such compositions typically comprise the nucleic acid molecule, protein, modulator, or antibody and a pharmaceutically acceptable carrier.

The term "administer" is used in its broadest sense and includes any method of introducing the compositions of the present invention into a subject. This includes

producing polypeptides or polynucleotides *in vivo* by *in vivo* transcription or translation of polynucleotides that have been exogenously introduced into a subject. Thus, polypeptides or nucleic acids produced in the subject from the exogenous compositions are encompassed in the term "administer."

As used herein the language "pharmaceutically acceptable carrier" is intended to include any and all solvents, dispersion media, coatings, antibacterial and antifungal agents, isotonic and absorption delaying agents, and the like, compatible with pharmaceutical administration. The use of such media and agents for pharmaceutically active substances is well known in the art. Except insofar as any conventional media or agent is incompatible with the active compound, such media can be used in the compositions of the invention. Supplementary active compounds can also be incorporated into the compositions. A pharmaceutical composition of the invention is formulated to be compatible with its intended route of administration. Examples of routes of administration include parenteral, e.g., intravenous, intradermal, subcutaneous, oral (e.g., inhalation), transdermal (topical), transmucosal, and rectal administration. Solutions or suspensions used for parenteral, intradermal, or subcutaneous application can include the following components: a sterile diluent such as water for injection, saline solution, fixed oils, polyethylene glycols, glycerine, propylene glycol or other synthetic solvents; antibacterial agents such as benzyl alcohol or methyl parabens; antioxidants such as ascorbic acid or sodium bisulfite; chelating agents such as ethylenediaminetetraacetic acid; buffers such as acetates, citrates or phosphates and agents for the adjustment of tonicity such as sodium chloride or dextrose. pH can be adjusted with acids or bases, such as hydrochloric acid or sodium hydroxide. The parenteral preparation can be enclosed in ampules, disposable syringes or multiple dose vials made of glass or plastic.

Pharmaceutical compositions suitable for injectable use include sterile aqueous solutions (where water soluble) or dispersions and sterile powders for the extemporaneous preparation of sterile injectable solutions or dispersion. For intravenous administration, suitable carriers include physiological saline, bacteriostatic water, Cremophor EL™ (BASF, Parsippany, NJ) or phosphate buffered saline (PBS). In all cases, the composition must be sterile and should be fluid to the extent that easy syringability exists. It must be stable under the conditions of manufacture and storage

and must be preserved against the contaminating action of microorganisms such as bacteria and fungi. The carrier can be a solvent or dispersion medium containing, for example, water, ethanol, polyol (for example, glycerol, propylene glycol, and liquid polyethylene glycol, and the like), and suitable mixtures thereof. The proper fluidity can be maintained, for example, by the use of a coating such as lecithin, by the maintenance of the required particle size in the case of dispersion and by the use of surfactants. Prevention of the action of microorganisms can be achieved by various antibacterial and antifungal agents, for example, parabens, chlorobutanol, phenol, ascorbic acid, thimerosal, and the like. In many cases, it will be preferable to include isotonic agents, for example, sugars, polyalcohols such as mannitol, sorbitol, sodium chloride in the composition. Prolonged absorption of the injectable compositions can be brought about by including in the composition an agent which delays absorption, for example, aluminum monostearate and gelatin.

Sterile injectable solutions can be prepared by incorporating the active compound (e.g., a sulfatase protein or anti-sulfatase antibody) in the required amount in an appropriate solvent with one or a combination of ingredients enumerated above, as required, followed by filtered sterilization. Generally, dispersions are prepared by incorporating the active compound into a sterile vehicle which contains a basic dispersion medium and the required other ingredients from those enumerated above. In the case of sterile powders for the preparation of sterile injectable solutions, the preferred methods of preparation are vacuum drying and freeze-drying which yields a powder of the active ingredient plus any additional desired ingredient from a previously sterile-filtered solution thereof.

Oral compositions generally include an inert diluent or an edible carrier. They can be enclosed in gelatin capsules or compressed into tablets. For oral administration, the agent can be contained in enteric forms to survive the stomach or further coated or mixed to be released in a particular region of the GI tract by known methods. For the purpose of oral therapeutic administration, the active compound can be incorporated with excipients and used in the form of tablets, troches, or capsules. Oral compositions can also be prepared using a fluid carrier for use as a mouthwash, wherein the compound in the fluid carrier is applied orally and swished and expectorated or swallowed. Pharmaceutically compatible binding agents, and/or adjuvant materials can be included

as part of the composition. The tablets, pills, capsules, troches and the like can contain any of the following ingredients, or compounds of a similar nature: a binder such as microcrystalline cellulose, gum tragacanth or gelatin; an excipient such as starch or lactose, a disintegrating agent such as alginic acid, Primogel, or corn starch; a lubricant such as magnesium stearate or Sterotes; a glidant such as colloidal silicon dioxide; a sweetening agent such as sucrose or saccharin; or a flavoring agent such as peppermint, methyl salicylate, or orange flavoring.

For administration by inhalation, the compounds are delivered in the form of an aerosol spray from pressured container or dispenser, which contains a suitable propellant, e.g., a gas such as carbon dioxide, or a nebulizer.

Systemic administration can also be by transmucosal or transdermal means. For transmucosal or transdermal administration, penetrants appropriate to the barrier to be permeated are used in the formulation. Such penetrants are generally known in the art, and include, for example, for transmucosal administration, detergents, bile salts, and fusidic acid derivatives. Transmucosal administration can be accomplished through the use of nasal sprays or suppositories. For transdermal administration, the active compounds are formulated into ointments, salves, gels, or creams as generally known in the art.

The compounds can also be prepared in the form of suppositories (e.g., with conventional suppository bases such as cocoa butter and other glycerides) or retention enemas for rectal delivery.

In one embodiment, the active compounds are prepared with carriers that will protect the compound against rapid elimination from the body, such as a controlled release formulation, including implants and microencapsulated delivery systems. Biodegradable, biocompatible polymers can be used, such as ethylene vinyl acetate, polyanhydrides, polyglycolic acid, collagen, polyorthoesters, and polylactic acid. Methods for preparation of such formulations will be apparent to those skilled in the art. The materials can also be obtained commercially from Alza Corporation and Nova Pharmaceuticals, Inc. Liposomal suspensions (including liposomes targeted to infected cells with monoclonal antibodies to viral antigens) can also be used as pharmaceutically acceptable carriers. These can be prepared according to methods known to those skilled in the art, for example, as described in U.S. Patent No. 4,522,811.

It is especially advantageous to formulate oral or parenteral compositions in dosage unit form for ease of administration and uniformity of dosage. "Dosage unit form" as used herein refers to physically discrete units suited as unitary dosages for the subject to be treated; each unit containing a predetermined quantity of active compound calculated to produce the desired therapeutic effect in association with the required pharmaceutical carrier. The specification for the dosage unit forms of the invention are dictated by and directly dependent on the unique characteristics of the active compound and the particular therapeutic effect to be achieved, and the limitations inherent in the art of compounding such an active compound for the treatment of individuals.

The nucleic acid molecules of the invention can be inserted into vectors and used as gene therapy vectors. Gene therapy vectors can be delivered to a subject by, for example, intravenous injection, local administration (U.S. 5,328,470) or by stereotactic injection (see e.g., Chen *et al.* (1994) *PNAS 91*:3054-3057). The pharmaceutical preparation of the gene therapy vector can include the gene therapy vector in an acceptable diluent, or can comprise a slow release matrix in which the gene delivery vehicle is imbedded. Alternatively, where the complete gene delivery vector can be produced intact from recombinant cells, e.g. retroviral vectors, the pharmaceutical preparation can include one or more cells which produce the gene delivery system.

As defined herein, a therapeutically effective amount of protein or polypeptide (i.e., an effective dosage) ranges from about 0.001 to 30 mg/kg body weight, preferably about 0.01 to 25 mg/kg body weight, more preferably about 0.1 to 20 mg/kg body weight, and even more preferably about 1 to 10 mg/kg, 2 to 9 mg/kg, 3 to 8 mg/kg, 4 to 7 mg/kg, or 5 to 6 mg/kg body weight.

The skilled artisan will appreciate that certain factors may influence the dosage required to effectively treat a subject, including but not limited to the severity of the disease or disorder, previous treatments, the general health and/or age of the subject, and other diseases present. Moreover, treatment of a subject with a therapeutically effective amount of a protein, polypeptide, or antibody can include a single treatment or, preferably, can include a series of treatments. In a preferred example, a subject is treated with antibody, protein, or polypeptide in the range of between about 0.1 to 20 mg/kg body weight, one time per week for between about 1 to 10 weeks, preferably between 2 to 8 weeks, more preferably between about 3 to 7

weeks, and even more preferably for about 4, 5, or 6 weeks. It will also be appreciated that the effective dosage of antibody, protein, or polypeptide used for treatment may increase or decrease over the course of a particular treatment. Changes in dosage may result and become apparent from the results of diagnostic assays as described herein.

The present invention encompasses agents which modulate expression or activity. An agent may, for example, be a small molecule. For example, such small molecules include, but are not limited to, peptides, peptidomimetics, amino acids, amino acid analogs, polynucleotides, polynucleotide analogs, nucleotides, nucleotide analogs, organic or inorganic compounds (i.e., including heteroorganic and organometallic compounds) having a molecular weight less than about 10,000 grams per mole, organic or inorganic compounds having a molecular weight less than about 5,000 grams per mole, organic or inorganic compounds having a molecular weight less than about 1,000 grams per mole, organic or inorganic compounds having a molecular weight less than about 500 grams per mole, and salts, esters, and other pharmaceutically acceptable forms of such compounds.

It is understood that appropriate doses of small molecule agents depends upon a number of factors within the ken of the ordinarily skilled physician, veterinarian, or researcher. The dose(s) of the small molecule will vary, for example, depending upon the identity, size, and condition of the subject or sample being treated, further depending upon the route by which the composition is to be administered, if applicable, and the effect which the practitioner desires the small molecule to have upon the nucleic acid or polypeptide of the invention. Exemplary doses include milligram or microgram amounts of the small molecule per kilogram of subject or sample weight (e.g., about 1 microgram per kilogram to about 500 milligrams per kilogram, about 100 micrograms per kilogram to about 5 milligrams per kilogram, or about 1 microgram per kilogram to about 50 micrograms per kilogram. It is furthermore understood that appropriate doses of a small molecule depend upon the potency of the small molecule with respect to the expression or activity to be modulated. Such appropriate doses may be determined using the assays described herein. When one or more of these small molecules is to be administered to an animal (e.g., a human) in order to modulate expression or activity of a polypeptide or nucleic

acid of the invention, a physician, veterinarian, or researcher may, for example, prescribe a relatively low dose at first, subsequently increasing the dose until an appropriate response is obtained. In addition, it is understood that the specific dose level for any particular animal subject will depend upon a variety of factors including the activity of the specific compound employed, the age, body weight, general health, gender, and diet of the subject, the time of administration, the route of administration, the rate of excretion, any drug combination, and the degree of expression or activity to be modulated.

The pharmaceutical compositions can be included in a container, pack, or dispenser together with instructions for administration.

Other Embodiments

In another aspect, the invention features, a method of analyzing a plurality of capture probes. The method can be used, e.g., to analyze gene expression. The method includes: providing a two dimensional array having a plurality of addresses, each address of the plurality being positionally distinguishable from each other address of the plurality, and each address of the plurality having a unique capture probe, e.g., a nucleic acid or peptide sequence; contacting the array with a 22438, 23553, 25278, or 26212 nucleic acid, preferably purified, polypeptide, preferably purified, or antibody, and thereby evaluating the plurality of capture probes. Binding, e.g., in the case of a nucleic acid, hybridization with a capture probe at an address of the plurality, is detected, e.g., by signal generated from a label attached to the 22438, 23553, 25278, or 26212 nucleic acid, polypeptide, or antibody.

The capture probes can be a set of nucleic acids from a selected sample, e.g., a sample of nucleic acids derived from a control or non-stimulated tissue or cell.

The method can include contacting the 22438, 23553, 25278, or 26212 nucleic acid, polypeptide, or antibody with a first array having a plurality of capture probes and a second array having a different plurality of capture probes. The results of each hybridization can be compared, e.g., to analyze differences in expression between a first and second sample. The first plurality of capture probes can be from a control sample, e.g., a wild type, normal, or non-diseased, non-stimulated, sample, e.g., a biological fluid, tissue, or cell sample. The second plurality of capture probes can be

from an experimental sample, e.g., a mutant type, at risk, disease-state or disorder-state, or stimulated, sample, e.g., a biological fluid, tissue, or cell sample.

The plurality of capture probes can be a plurality of nucleic acid probes each of which specifically hybridizes with an allele of 22438, 23553, 25278, or 26212. Such methods can be used to diagnose a subject, e.g., to evaluate risk for a disease or disorder, to evaluate suitability of a selected treatment for a subject, to evaluate whether a subject has a disease or disorder. 22438, 23553, 25278, or 26212 are associated with sulfatase activity, thus it is useful for disorders associated with abnormal sulfatase activity.

The method can be used to detect SNPs, as described below.

In another aspect, the invention features, a method of analyzing a plurality of probes. The method is useful, e.g., for analyzing gene expression. The method includes: providing a two dimensional array having a plurality of addresses, each address of the plurality being positionally distinguishable from each other address of the plurality having a unique capture probe, e.g., wherein the capture probes are from a cell or subject which express or misexpress 22438, 23553, 25278, or 26212, or from a cell or subject in which a 22438, 23553, 25278, or 26212 mediated response has been elicited, e.g., by contact of the cell with 22438, 23553, 25278, or 26212 nucleic acid or protein, or administration to the cell or subject 22438, 23553, 25278, or 26212 nucleic acid or protein; contacting the array with one or more inquiry probe, wherein an inquiry probe can be a nucleic acid, polypeptide, or antibody (which is preferably other than 22438, 23553, 25278, or 26212 nucleic acid, polypeptide, or antibody); providing a two dimensional array having a plurality of addresses, each address of the plurality being positionally distinguishable from each other address of the plurality, and each address of the plurality having a unique capture probe, e.g., wherein the capture probes are from a cell or subject which does not express 22438, 23553, 25278, or 26212 (or does not express as highly as in the case of the 22438, 23553, 25278, or 26212 positive plurality of capture probes) or from a cell or subject which in which a 22438, 23553, 25278, or 26212 mediated response has not been elicited (or has been elicited to a lesser extent than in the first sample); contacting the array with one or more inquiry probes (which is preferably other than a 22438, 23553, 25278, or 26212 nucleic acid, polypeptide, or antibody), and thereby evaluating the plurality of

capture probes. Binding, e.g., in the case of a nucleic acid, hybridization with a capture probe at an address of the plurality, is detected, e.g., by signal generated from a label attached to the nucleic acid, polypeptide, or antibody.

In another aspect, the invention features a method of analyzing 22438, 23553, 25278, or 26212, e.g., analyzing structure, function, or relatedness to other nucleic acid or amino acid sequences. The method includes: providing a 22438, 23553, 25278, or 26212 nucleic acid or amino acid sequence; comparing the 22438, 23553, 25278, or 26212 sequence with one or more preferably a plurality of sequences from a collection of sequences, e.g., a nucleic acid or protein sequence database; to thereby analyze 22438, 23553, 25278, or 26212.

Preferred databases include GenBank™. The method can include evaluating the sequence identity between a 22438, 23553, 25278, or 26212 sequence and a database sequence. The method can be performed by accessing the database at a second site, e.g., over the internet.

In another aspect, the invention features, a set of oligonucleotides, useful, e.g., for identifying SNP's, or identifying specific alleles of 22438, 23553, 25278, or 26212. The set includes a plurality of oligonucleotides, each of which has a different nucleotide at an interrogation position, e.g., an SNP or the site of a mutation. In a preferred embodiment, the oligonucleotides of the plurality identical in sequence with one another (except for differences in length). The oligonucleotides can be provided with different labels, such that an oligonucleotides which hybridizes to one allele provides a signal that is distinguishable from an oligonucleotides which hybridizes to a second allele.

This invention is further illustrated by the following examples which should not be construed as limiting. The contents of all references, patents and published patent applications cited throughout this application are incorporated herein by reference.

EXAMPLES

Example 1: Identification and Characterization of Human 22438 cDNAs

The human 22438 sequence (Figure 1A-B; SEQ ID NO:2), which is approximately 2175 nucleotides long including untranslated regions, contains a predicted methionine-initiated coding sequence of about 1578 nucleotides (nucleotides 248-1825 of SEQ ID NO:2; SEQ ID NO:11). The coding sequence encodes a 525 amino acid protein (SEQ ID NO:1).

PFAM analysis indicates that 22438 contains a sulfatase domain. For general information regarding PFAM identifiers, PS prefix and PF prefix domain identification numbers, refer to Sonnhammer *et al.* (1997) *Protein* 28:405-420 and http//www.psc.edu/general/software/packages/pfam/pfam.html.

As used herein, the term "sulfatase domain" includes an amino acid sequence of about 80-420 amino acid residues in length and having a bit score for the alignment of the sequence to the sulfatase domain (HMM) of at least 8. Preferably, a sulfatase domain includes at least about 100-250 amino acids, more preferably about 130-200 amino acid residues, or about 160-200 amino acids and has a bit score for the alignment of the sequence to the sulfatase domain (HMM) of at least 16 or greater. The sulfatase domain (HMM) has been assigned the PFAM Accession PF00884 (http;//pfam.wustl.edu/). An alignment of the sulfatase domain (amino acids 36-462 of SEQ ID NO:1) of human 22438 with a consensus amino acid sequence derived from a hidden Markov model is depicted in Figure 19.

In a preferred embodiment 22438-like polypeptide or protein has a "sulfatase domain" or a region which includes at least about 100-250, more preferably about 130-200 or 160-200, amino acid residues and has at least about 60%, 70%, 80%, 90%, 95%, 99%, or 100% sequence identity with a "sulfatase domain," e.g., the sulfatase domain of human 22438-like polypeptide or protein (e.g., amino acid residues 36-462 of SEQ ID NO:1).

To identify the presence of an "sulfatase" domain in a 22438-like protein sequence, and make the determination that a polypeptide or protein of interest has a

particular profile, the amino acid sequence of the protein can be searched against a database of HMMs (e.g., the Pfam database, release 2.1) using the default parameters (http://www.sanger.ac.uk/Software/Pfam/HMM_search). For example, the hmmsf program, which is available as part of the HMMER package of search programs, is a family specific default program for MILPAT0063 and a score of 15 is the default threshold score for determining a hit. Alternatively, the threshold score for determining a hit can be lowered (e.g., to 8 bits). A description of the Pfam database can be found in Sonhammer *et al.* (1997) *Proteins 28(3)*:405-420 and a detailed description of HMMs can be found, for example, in Gribskov *et al.* (1990) *Meth. Enzymol. 183*:146-159; Gribskov *et al.* (1987) *Proc. Natl. Acad. Sci. USA 84*:4355-4358; Krogh *et al.* (1994) *J. Mol. Biol. 235*:1501-1531; and Stultz *et al.* (1993) *Protein Sci. 2*:305-314, the contents of which are incorporated herein by reference.

Example 2: Tissue Distribution of 22348 mRNA

Northern blot hybridizations with various RNA samples are performed under standard conditions and washed under stringent conditions, i.e., 0.2 X SSC at 65°C. A DNA probe corresponding to all or a portion of the 22348 cDNA (SEQ ID NO:2) can be used. The DNA is radioactively labeled with ^{32}P-dCTP using the Prime-It Kit (Stratagene, La Jolla, CA) according to the instructions of the supplier. Filters containing mRNA from mouse hematopoietic and endocrine tissues, and cancer cell lines (Clontech, Palo Alto, CA) are probed in ExpressHyb hybridization solution (Clontech) and washed at high stringency according to manufacturer's recommendations.

Example 3: Identification and Characterization of Human 23553 cDNAs

The human 23553 sequence (Figure 5A-B; SEQ ID NO:4), which is approximately 4321 nucleotides long including untranslated regions, contains a predicted methionine-initiated coding sequence of about 2616 nucleotides (nucleotides 510-3125 of SEQ ID NO:4; SEQ ID NO:12). The coding sequence encodes a 871 amino acid protein (SEQ ID NO:3).

PFAM analysis indicates that 23553 has a sulfatase domain. For general information regarding PFAM identifiers, PS prefix and PF prefix domain

identification numbers, refer to Sonnhammer *et al.* (1997) *Protein* 28:405-420 and http//www.psc.edu/general/software/packages/pfam/pfam.html. An alignment of the sulfatase domain (amino acids 43 to 467 of SEQ ID NO:3) of human 23553-like with a consensus amino acid sequence derived from a hidden Markov model is depicted in Figure 20. For further information on sulfatase domains, see Example 1.

In one embodiment, a 23553-like protein includes at least one transmembrane domain. As used herein, the term "transmembrane domain" includes an amino acid sequence of about 15 amino acid residues in length that spans a phospholipid membrane. More preferably, a transmembrane domain includes about at least 18, 20, 22, or 24 amino acid residues and spans a phospholipid membrane. Transmembrane domains are rich in hydrophobic residues, and typically have an α-helical structure. In a preferred embodiment, at least 50%, 60%, 70%, 80%, 90%, 95% or more of the amino acids of a transmembrane domain are hydrophobic, e.g., leucines, isoleucines, tyrosines, or tryptophans. Transmembrane domains are described in, for example, http://pfam.wustl.edu/cgi-bin/getdesc?name=7tm-1, and Zagotta W.N. *et al.* (1996) *Annual Rev. Neuronsci. 19*:235-63, the contents of which are incorporated herein by reference.

In a preferred embodiment, a 23553-like polypeptide or protein has at least one transmembrane domain or a region which includes at least 18, 20, 22, or 24 amino acid residues and has at least about 60%, 70% 80% 90% 95%, 99%, or 100% sequence identity with a "transmembrane domain," e.g., at least one transmembrane domain of human 23553 (e.g., amino acid residues 7 to 25 of SEQ ID NO:3).

In another embodiment, a 23553 protein includes at least one "non-transmembrane domain." As used herein, "non-transmembrane domains" are domains that reside outside of the membrane. When referring to plasma membranes, non-transmembrane domains include extracellular domains (i.e., outside of the cell) and intracellular domains (i.e., within the cell). When referring to membrane-bound proteins found in intracellular organelles (e.g., mitochondria, endoplasmic reticulum, peroxisomes and microsomes), non-transmembrane domains include those domains of the protein that reside in the cytosol (i.e., the cytoplasm), the lumen of the organelle, or the matrix or the intermembrane space (the latter two relate specifically to mitochondria organelles). The C-terminal amino acid residue of a non-transmembrane

domain is adjacent to an N-terminal amino acid residue of a transmembrane domain in a naturally occurring 23553-like protein.

In a preferred embodiment, a 23553-like polypeptide or protein has a "non-transmembrane domain" or a region which includes at least about 1-350, preferably about 200-320, more preferably about 230-300, and even more preferably about 240-280 amino acid residues, and has at least about 60%, 70% 80% 90% 95%, 99% or 100% sequence identity with a "non-transmembrane domain", e.g., a non-transmembrane domain of human 23553-like protein.

A non-transmembrane domain located at the N-terminus of a 23553-like protein or polypeptide is referred to herein as an "N-terminal non-transmembrane domain." As used herein, an "N-terminal non-transmembrane domain" includes an amino acid sequence having about 1-100. For example, an N-terminal non-transmembrane domain is located at about amino acid residues 1 to 6 of SEQ ID NO:3.

Similarly, a non-transmembrane domain located at the C-terminus of a 23553-like protein or polypeptide is referred to herein as a "C-terminal non-transmembrane domain." As used herein, a "C-terminal non-transmembrane domain" includes an amino acid sequence having about 1-800, preferably about 15-500, preferably about 20-270, more preferably about 25-255 amino acid residues in length and is located outside the boundaries of a membrane. For example, a C-terminal non-transmembrane domain is located at about amino acid residues 26-871 of SEQ ID NO:3.

The ORF analyzer predicts that 23553 has a signal peptide. Therefore, a 23553-like molecule can further include a signal sequence. As used herein, a "signal sequence" refers to a peptide of about 20-80 amino acid residues in length which occurs at the N-terminus of secretory and integral membrane proteins and which contains a majority of hydrophobic amino acid residues. For example, a signal sequence contains at least about 12-25 amino acid residues, preferably about 30-70 amino acid residues, and has at least about 40-70%, preferably about 50-65%, and more preferably about 55-60% hydrophobic amino acid residues (e.g., alanine, valine, leucine, isoleucine, phenylalanine, tyrosine, tryptophan, or proline). Such a "signal sequence", also referred to in the art as a "signal peptide", serves to direct a protein

containing such a sequence to a lipid bilayer. For example, in one embodiment, a 23553-like protein contains a signal sequence of about amino acids 1-22 of SEQ ID NO:3. The "signal sequence" is cleaved during processing of the mature protein. The mature 23553-like protein corresponds to amino acids 23-871 of SEQ ID NO:3.

CLUSTAL multiple sequence alignment analysis shows homology between 23553 and the following sequences (identified by GenBank accession number): P14217, *Chlamydomonas reinhardtii* arylsulfatase; Q10723, *Volvox carteri* arylsulfatase; CAB40661, human N-acetylglucosamine-6-sulfatase homolog; P15586, human N-acetylglucosamine-6-sulfatase; P50426, goat N-acetylglucosamine-6-sulfatase; AAA83618, *C. elegans* putative sulfatase; AAC02716, *Neurospora crassa* arylsulfatase; P31447, *E. coli* hypothetical sulfatase.

Example 4: Tissue Distribution of 23553 mRNA

In normal human tissues tested, high expression of 23553 was observed in trachea, vein, osteoblast, kidney, and testes. Significant expression of 23553 was found in adipose, colon, skeletal muscle, thyroid, prostate, and other tissues. *See* Figure 25. In comparisons of normal and tumor tissue, 23553 expression was detected in all samples tested, with increased expression in breast, colon, and lung tumors. *See* Figure 26. Further, elevated expression of 23553 was found in glioblastoma samples, as compared to normal brain tissue samples. Expression levels were determined by quantitative PCR (Taqman® brand quantitative PCR kit, Applied Biosystems). The quantitative PCR reactions were performed according to the kit manufacturer's instructions.

cDNA library array analysis of 23553 revealed expression in adipose, adrenal gland, bone, brain, colon, colon metastases to liver, endothelial, heart, liver, lung, muscle, osteoblast, skin, testes, thyroid, and other tissue. Reverse transcriptase polymerase chain reaction (RT-PCR) revealed 23553 expression in clinical samples of normal and tumor colon tissue, normal and metastatic liver tissue, and in lung squamous cell carcinoma tissue. *In situ* hybridization showed expression of 23553 in the following tissues: 3 of 3 breast tumor; 0 of 2 normal breast; 4 of 4 lung tumor; 0 of 2 normal lung; 4 of 4 colon tumor; and 2 of 2 liver metasteses. In all cases,

expression of 23553 was confined to the stromal component of tissue; no expression was detected in normal or tumor epithelium.

Angiogenic growth factors (e.g., bFGF) are present in the extracellular matrix (ECM), and can be released from the ECM by heparinase-like enzymes. This includes the glycosyl-sulfatases. The released growth factors in turn stimulate blood vessel formation. *See* Baird A, Ling N., "Fibroblast growth factors are present in the extracellular matrix produced by endothelial cells in vitro: implications for a role of heparinase-like enzymes in the neovascular response," *Biochem Biophys Res Commun.* (1987) 142(2):428-35.

As noted, 23553 has amino acid sequence features that place it in the class of glycosyl sulfate cleaving enzymes. Taqman results (above) show that its expression is elevated in clinical tumor samples. *In situ* hybridization shows specific, localized 23553 expression in the tumor stromal component of all tumor samples tested, whereas its expression is low or absent in normal tissues. This suggests that, through catalytic activity, 23553 promotes tumor growth or is involved in tumor maintenance by degrading the ECM and releasing growth factors.

Example 5: Identification and Characterization of Human 25278 cDNAs

The human 25278 sequence (Figure 10A-B; SEQ ID NO:6), which is approximately 2940 nucleotides long including untranslated regions, contains a predicted methionine-initiated coding sequence of about 1710 nucleotides (nucleotides 334-2043 of SEQ ID NO:6; SEQ ID NO:13). The coding sequence encodes a 569 amino acid protein (SEQ ID NO:5).

PFAM analysis indicates that 25278 has a sulfatase domain. For general information regarding PFAM identifiers, PS prefix and PF prefix domain identification numbers, refer to Sonnhammer *et al.* (1997) *Protein* 28:405-420 and http//www.psc.edu/general/software/packages/pfam/pfam.html. An alignment of the sulfatase domain (amino acids 47 to 471 of SEQ ID NO:5) of human 25278 with a consensus amino acid sequence derived from a hidden Markov model is depicted in Figure 27. For further information on sulfatase domains, see Example 1.

Example 6: Identification and Characterization of Human 26212 cDNAs

The human 26212 sequence (Figure 15; SEQ ID NO:8), which is approximately 2253 nucleotides long including untranslated regions, contains a predicted methionine-initiated coding sequence of about 1800 nucleotides (nucleotides 324-2123 of SEQ ID NO:8; SEQ ID NO:14). The coding sequence encodes a 599 amino acid protein (SEQ ID NO:7).

PFAM analysis indicates that 26212 has a sulfatase domain. For general information regarding PFAM identifiers, PS prefix and PF prefix domain identification numbers, refer to Sonnhammer *et al.* (1997) *Protein* 28:405-420 and http//www.psc.edu/general/software/packages/pfam/pfam.html. An alignment of the sulfatase domain (amino acids 76-502 of SEQ ID NO:7) of human 26212 with a consensus amino acid sequence derived from a hidden Markov model is depicted in Figure 29. For further information on sulfatase domains, see Example 1.

In one embodiment, 26212-like protein includes at least one transmembrane domain. As used herein, the term "transmembrane domain" includes an amino acid sequence of about 15 amino acid residues in length that spans a phospholipid membrane. More preferably, a transmembrane domain includes about at least 18, 20, 22, or 24 amino acid residues and spans a phospholipid membrane. For more information on transmembrane domains, see example 3.

In a preferred embodiment, a 26212-like polypeptide or protein has at least one transmembrane domain or a region which includes at least 18, 20, 22, 24, 25, or 30 amino acid residues and has at least about 60%, 70% 80% 90% 95%, 99%, or 100% sequence identity with a "transmembrane domain," e.g., at least one transmembrane domain of human 26212-like polypeptide or protein (e.g., amino acid residues 24 to 44 of SEQ ID NO:7).

In another embodiment, a 26212-like protein includes at least one "non-transmembrane domain." The C-terminal amino acid residue of a non-transmembrane domain is adjacent to an N-terminal amino acid residue of a transmembrane domain in a naturally occurring 26212-like protein. For more information on non-transmembrane domains, see Example 3.

In a preferred embodiment, a 26212-like polypeptide or protein has a "non-transmembrane domain" or a region which includes at least about 1-350, preferably about 200-320, more preferably about 230-300, and even more preferably about 240-280 amino acid residues, and has at least about 60%, 70% 80% 90% 95%, 99% or 100% sequence identity with a "non-transmembrane domain", e.g., a non-transmembrane domain of human 26212-like polypeptide or protein. An N-terminal non-transmembrane domain is located at about amino acid residues 1 to 23 of SEQ ID NO:7. A C-terminal non-transmembrane domain is located at about amino acid residues 45 to 599 of SEQ ID NO:7.

A 26212-like molecule can further include a signal sequence. For more information on signal sequences, see Example 3.

Example 7: Tissue Distribution of 26212 mRNA

In six independent experiments, 26212 showed higher levels of expression in proliferating endothelial cells as compared to arrested endothelial cells. 26212 expression was also higher in proliferating endothelial cells than in non-endothelial cells. *See* Figure 30. 26212 expression levels were upregulated in breast tissue cell lines treated with epidermal growth factor, as well. *See* Figure 34. 26212 is expressed in hemangiomas and other angiogenic tissues, including fetal heart, uterine adenocarcinoma, and endometrial polyps. *See* Figure 35. Endothelial and glial cells showed higher levels of 26212 expression as compared to other tissues and cells. *See* Figure 36. 26212 also showed higher levels of expressing in some lung, breast and brain tumors as compared to normal tissues. Expression levels of 26212 were found to be higher in proliferating endothelial cells than in tumors, too. Expression levels were determined by quantitative PCR (Taqman® brand quantitative PCR kit, Applied Biosystems). The quantitative PCR reactions were performed according to the kit manufacturer's instructions.

In situ hybridization analysis was also carried out. 26212 showed weak expression in ovarian tumor, and no expression in normal ovary. Similarly, colon metastases showed weak expression of 26212, and normal colon tissue and primary tumors showed no expression. A subset of lung tumors tested showed expression of 26212, while no expression was revealed in normal lung.

Angiogenic growth factors (e.g., bFGF) are present in the extracellular matrix (ECM), and can be released from the ECM by heparinase-like enzymes. This includes the glycosyl-sulfatases. The released growth factors in turn stimulate blood vessel formation by, e.g., attracting endothelial cells to form new vessels. *See* Baird A, Ling N., "Fibroblast growth factors are present in the extracellular matrix produced by endothelial cells in vitro: implications for a role of heparinase-like enzymes in the neovascular response," *Biochem Biophys Res Commun.* (1987) 142(2):428-35.

As noted, 26212 has amino acid sequence features that place it in the class of glycosyl sulfate cleaving enzymes. Taqman results (above) show that its expression is elevated in proliferating endothelial cells, suggesting that 26212 is specifically involved in active angiogenic sites.

Example 8: Recombinant Expression of 22348, 23553, 25278, or 26212 in Bacterial Cells

In this example, 22348, 23553, 25278, or 26212 is expressed as a recombinant glutathione-S-transferase (GST) fusion polypeptide in *E. coli* and the fusion polypeptide is isolated and characterized. Specifically, 22348, 23553, 25278, or 26212 is fused to GST and this fusion polypeptide is expressed in *E. coli*, e.g., strain PEB199. Expression of the GST-26212 fusion protein in PEB199 is induced with IPTG. The recombinant fusion polypeptide is purified from crude bacterial lysates of the induced PEB199 strain by affinity chromatography on glutathione beads. Using polyacrylamide gel electrophoretic analysis of the polypeptide purified from the bacterial lysates, the molecular weight of the resultant fusion polypeptide is determined.

Example 9: Expression of Recombinant 22348, 23553, 25278, or 26212 Protein in COS Cells

To express the 22348, 23553, 25278, or 26212 gene in COS cells, the pcDNA/Amp vector by Invitrogen Corporation (San Diego, CA) is used. This vector contains an SV40 origin of replication, an ampicillin resistance gene, an *E. coli* replication origin, a CMV promoter followed by a polylinker region, and an SV40 intron and polyadenylation site. A DNA fragment encoding the entire 22348, 23553,

25278, or 26212 protein and an HA tag (Wilson *et al.* (1984) *Cell* 37:767) or a FLAG tag fused in-frame to its 3' end of the fragment is cloned into the polylinker region of the vector, thereby placing the expression of the recombinant protein under the control of the CMV promoter.

To construct the plasmid, the 22348, 23553, 25278, or 26212 DNA sequence is amplified by PCR using two primers. The 5' primer contains the restriction site of interest followed by approximately twenty nucleotides of the 22348, 23553, 25278, or 26212 coding sequence starting from the initiation codon; the 3' end sequence contains complementary sequences to the other restriction site of interest, a translation stop codon, the HA tag or FLAG tag and the last 20 nucleotides of the 22348, 23553, 25278, or 26212 coding sequence. The PCR amplified fragment and the pCDNA/Amp vector are digested with the appropriate restriction enzymes and the vector is dephosphorylated using the CIAP enzyme (New England Biolabs, Beverly, MA). Preferably the two restriction sites chosen are different so that the 22348, 23553, 25278, or 26212 gene is inserted in the correct orientation. The ligation mixture is transformed into *E. coli* cells (strains HB101, DH5α, SURE, available from Stratagene Cloning Systems, La Jolla, CA, can be used), the transformed culture is plated on ampicillin media plates, and resistant colonies are selected. Plasmid DNA is isolated from transformants and examined by restriction analysis for the presence of the correct fragment.

COS cells are subsequently transfected with the 22348, 23553, 25278, or 26212-pcDNA/Amp plasmid DNA using the calcium phosphate or calcium chloride co-precipitation methods, DEAE-dextran-mediated transfection, lipofection, or electroporation. Other suitable methods for transfecting host cells can be found in Sambrook, J., Fritsh, E. F., and Maniatis, T. *Molecular Cloning: A Laboratory Manual. 2nd, ed., Cold Spring Harbor Laboratory,* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1989. The expression of the 22348, 23553, 25278, or 26212 polypeptide is detected by radiolabelling (^{35}S-methionine or ^{35}S-cysteine available from NEN, Boston, MA, can be used) and immunoprecipitation (Harlow, E. and Lane, D. *Antibodies: A Laboratory Manual*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1988) using an HA specific monoclonal antibody. Briefly, the cells are labeled for 8 hours with ^{35}S-methionine (or ^{35}S-cysteine). The

culture media are then collected and the cells are lysed using detergents (RIPA buffer, 150 mM NaCl, 1% NP-40, 0.1% SDS, 0.5% DOC, 50 mM Tris, pH 7.5). Both the cell lysate and the culture media are precipitated with an HA specific monoclonal antibody. Precipitated polypeptides are then analyzed by SDS-PAGE.

Alternatively, DNA containing the 22348, 23553, 25278, or 26212 coding sequence is cloned directly into the polylinker of the pCDNA/Amp vector using the appropriate restriction sites. The resulting plasmid is transfected into COS cells in the manner described above, and the expression of the 22348, 23553, 25278, or 26212 polypeptide is detected by radiolabelling and immunoprecipitation using a 22348, 23553, 25278, or 26212 specific monoclonal antibody.

This invention may be embodied in many different forms and should not be construed as limited to the embodiments set forth herein; rather, these embodiments are provided so that this disclosure will fully convey the invention to those skilled in the art. Many modifications and other embodiments of the invention will come to mind in one skilled in the art to which this invention pertains having the benefit of the teachings presented in the foregoing description. Although specific terms are employed, they are used as in the art unless otherwise indicated.

Applicant's or agent's file reference	35800/208709	International application No.	PCT/US01/

INDICATIONS RELATING TO DEPOSITED MICROORGANISM OR OTHER BIOLOGICAL MATERIAL

(PCT Rule 13*bis*)

A. The indications made below relate to the deposited microorganism or other biological material referred to in the description on page 5, line 31

B. IDENTIFICATION OF DEPOSIT Further deposits are identified on an additional sheet ☐

Name of depository institution
American Type Culture Collection

Address of depositary institution *(including postal code and country)*
10801 University Blvd.
Manassas, VA 20110-2209 US

Date of deposit	Accession Number
	PTA-

C. ADDITIONAL INDICATIONS *(leave blank if not applicable)* This information is continued on an additional sheet ☐

Page 17, line 12; page 22, line 9; page 23, line 23; page 108, lines 7, 13, 17, 21, 24 and 29; page 109, lines 8 and 13; page 110, lines 2, 6, 13 and 22; page 111, lines 1, 6, 9 and 13.

D. DESIGNATED STATES FOR WHICH INDICATIONS ARE MADE *(if the indicators are not for all designated States)*

E. SEPARATE FURNISHING OF INDICATIONS *(leave blank if not applicable)*

The indications listed below will be submitted to the International Bureau later *(specify the general nature of the indications e.g., "Accession Number of Deposit")*

Accession Number of Deposit and Date of Deposit

For receiving Office use only	For International Bureau use only
☑ This sheet was received with the international application	☐ This sheet was received with the International Bureau on:
Authorized officer MELVIN S. BROOKS SR INTERNATIONAL DIVISION 703-305-5163 MSB	Authorized officer

Form PCT/RO/134 (July 1998)

Applicant's or agent's file reference	35800/208709	International application No.	PCT/US01/

INDICATIONS RELATING TO DEPOSITED MICROORGANISM OR OTHER BIOLOGICAL MATERIAL

(PCT Rule 13*bis*)

A. The indications made below relate to the deposited microorganism or other biological material referred to in the description on page 5, line 31

B. IDENTIFICATION OF DEPOSIT Further deposits are identified on an additional sheet ☐

Name of depository institution
American Type Culture Collection

Address of depositary institution *(including postal code and country)*
10801 University Blvd.
Manassas, VA 20110-2209 US

Date of deposit	Accession Number
05 April 2000 (05.04.00)	PTA- 1639

C. ADDITIONAL INDICATIONS *(leave blank if not applicable)* This information is continued on an additional sheet ☐

Page 17, line 12; page 22, line 10; page 23, line 23; page 108, lines 7, 13, 17, 21, 24 and 29; page 109, lines 8 and 13; page 110, lines 2, 6, 13 and 22; page 111, lines 1, 6, 9 and 13.

D. DESIGNATED STATES FOR WHICH INDICATIONS ARE MADE *(if the indicators are not for all designated States)*

E. SEPARATE FURNISHING OF INDICATIONS *(leave blank if not applicable)*

The indications listed below will be submitted to the International Bureau later *(specify the general nature of the indications e.g., "Accession Number of Deposit")*

For receiving Office use only

☑ This sheet was received with the international application

Authorized officer MELVIN S. BROOKS SR
INTERNATIONAL DIVISION
703-305-5169
MSB

For International Bureau use only

☐ This sheet was received with the International Bureau on:

Authorized officer

Form PCT/RO/134 (July 1998)

Applicant's or agent's file reference	35800/208709	International application No.	PCT/US01/

INDICATIONS RELATING TO DEPOSITED MICROORGANISM OR OTHER BIOLOGICAL MATERIAL

(PCT Rule 13*bis*)

A. The indications made below relate to the deposited microorganism or other biological material referred to in the description on page 5, line 31

B. IDENTIFICATION OF DEPOSIT Further deposits are identified on an additional sheet ☐

Name of depository institution
American Type Culture Collection

Address of depositary institution *(including postal code and country)*
10801 University Blvd.
Manassas, VA 20110-2209 US

Date of deposit	Accession Number
09 May 2000 (09.05.00)	PTA- 1846

C. ADDITIONAL INDICATIONS *(leave blank if not applicable)* This information is continued on an additional sheet ☐

Page 17, line 12; page 22, line 10; page 23, line 23; page 108, lines 7, 13, 17, 21, 24 and 29; page 109, lines 8 and 13; page 110, lines 2, 6, 13 and 22; page 111, lines 1, 6, 9 and 13.

D. DESIGNATED STATES FOR WHICH INDICATIONS ARE MADE *(if the indicators are not for all designated States)*

E. SEPARATE FURNISHING OF INDICATIONS *(leave blank if not applicable)*

The indications listed below will be submitted to the International Bureau later *(specify the general nature of the indications e.g., "Accession Number of Deposit")*

For receiving Office use only	For International Bureau use only
☑ This sheet was received with the international application	☐ This sheet was received with the International Bureau on:
Authorized officer	Authorized officer

Form PCT/RO/134 (July 1998)

Applicant's or agent's file reference	35800/208709	International application No.	PCT/US01/

INDICATIONS RELATING TO DEPOSITED MICROORGANISM OR OTHER BIOLOGICAL MATERIAL

(PCT Rule 13*bis*)

A. The indications made below relate to the deposited microorganism or other biological material referred to in the description on page 5, line 32

B. IDENTIFICATION OF DEPOSIT Further deposits are identified on an additional sheet ☐

Name of depository institution
American Type Culture Collection

Address of depositary institution *(including postal code and country)*
10801 University Blvd.
Manassas, VA 20110-2209 US

Date of deposit	Accession Number
	PTA-

C. ADDITIONAL INDICATIONS *(leave blank if not applicable)* This information is continued on an additional sheet ☐

Page 17, line 12; page 22, line 10; page 23, line 23; page 108, lines 8, 13, 17, 21, 24 and 29; page 109, lines 9 and 13; page 110, lines 2, 6, 13 and 22; page 111, lines 2, 6, 9 and 13.

D. DESIGNATED STATES FOR WHICH INDICATIONS ARE MADE *(if the indicators are not for all designated States)*

E. SEPARATE FURNISHING OF INDICATIONS *(leave blank if not applicable)*

The indications listed below will be submitted to the International Bureau later *(specify the general nature of the indications e.g., "Accession Number of Deposit")*

Accession Number of Deposit and Date of Deposit

For receiving Office use only	For International Bureau use only
☑ This sheet was received with the international application	☐ This sheet was received with the International Bureau on:
Authorized officer MELVIN S. BROOKS SR. INTERNATIONAL DIVISION 703-305-5768	Authorized officer

THAT WHICH IS CLAIMED:

1. An isolated nucleic acid molecule selected from the group consisting of:

a) a nucleic acid molecule comprising a nucleotide sequence which is at least 60% identical to the nucleotide sequence of SEQ ID NO: 2, 4, 6, 8, 11, 12, 13, or 14, or the nucleotide sequence of the cDNA insert of the plasmid deposited with ATCC as Patent Deposit Number ____, PTA-1639, PTA-1846, or ____, wherein said nucleotide sequence encodes a polypeptide having biological activity;

b) a nucleic acid molecule comprising a fragment of at least 20 nucleotides of the nucleotide sequence of SEQ ID NO: 2, 4, 6, 8, 11, 12, 13, or 14, or the nucleotide sequence of the cDNA insert of the plasmid deposited with ATCC as Patent Deposit Number ____, PTA-1639, PTA-1846, or ____;

c) a nucleic acid molecule which encodes a polypeptide comprising the amino acid sequence of SEQ ID NO:1, 3, 5, or 7, or the amino acid sequence encoded by the cDNA insert of the plasmid deposited with the ATCC as Patent Deposit Number ____, PTA-1639, PTA-1846, or ____;

d) a nucleic acid molecule which encodes a fragment of a polypeptide comprising the amino acid sequence of SEQ ID NO:1, 3, 5, 7, or the amino acid sequence encoded by the cDNA insert of the plasmid deposited with the ATCC as Patent Deposit Number ____, PTA-1639, PTA-1846, or ____, wherein the fragment comprises at least 15 contiguous amino acids of SEQ ID NO:1, 3, 5, or 7, or the amino acid sequence encoded by the cDNA insert of the plasmid deposited with the ATCC as Patent Deposit Number ____, PTA-1639, PTA-1846, or ____;

e) a nucleic acid molecule which encodes a naturally occurring allelic variant of a biologically active polypeptide comprising the amino acid sequence of SEQ ID NO:1, 3, 5, or 7, or the amino acid sequence encoded by the cDNA insert of the plasmid deposited with the ATCC as Patent Deposit Number ____, PTA-1639, PTA-1846, or ____, wherein the nucleic acid molecule hybridizes to a nucleic acid molecule comprising the complement of SEQ ID NO:2, 4, 6, 8, 11, 12, 13, or 14 under stringent conditions; and

f) a nucleic acid molecule comprising the complement of a), b), c), d), or e).

2. The isolated nucleic acid molecule of claim 1, which is selected from the group consisting of:

a) a nucleic acid molecule comprising the nucleotide sequence of SEQ ID NO:2, 4, 6, 8, 11, 12, 13, 14, the cDNA insert of any one the plasmids deposited with ATCC as Patent Deposit Number _____, PTA-1639, PTA-1846, or _____, or a complement thereof; and

b) a nucleic acid molecule which encodes a polypeptide comprising the amino acid sequence of SEQ ID NO:1, 3, 5, or 7, or an amino acid sequence encoded by the cDNA insert of any of the plasmids deposited with ATCC as Patent Deposit Number _____, PTA-1639, PTA-1846, or _____.

3. The nucleic acid molecule of claim 1 further comprising vector nucleic acid sequences.

4. The nucleic acid molecule of claim 1 further comprising nucleic acid sequences encoding a heterologous polypeptide.

5. A host cell which contains the nucleic acid molecule of claim 1.

6. The host cell of claim 5 which is a mammalian host cell.

7. A nonhuman mammalian host cell containing the nucleic acid molecule of claim 1.

8. An isolated polypeptide selected from the group consisting of:

a) a biological active polypeptide which is encoded by a nucleic acid molecule comprising a nucleotide sequence which is at least 60% identical to a nucleic acid comprising the nucleotide sequence of SEQ ID NO: 2, 4, 6, 8, 11, 12, 13,

or 14 or the nucleotide sequence of the cDNA insert of the plasmid deposited with ATCC as Patent Deposit Number ____, PTA-1639, PTA-1846, or ____;

b) a naturally occurring allelic variant of a polypeptide comprising the amino acid sequence of SEQ ID NO:1, 3, 5, or 7, or the amino acid sequence encoded by the cDNA insert of the plasmid deposited with the ATCC as Patent Deposit Number ____, PTA-1639, PTA-1846, or ____, wherein the polypeptide is encoded by a nucleic acid molecule which hybridizes to a nucleic acid molecule comprising the complement of SEQ ID NO: 2, 4, 6, 8, 11, 12, 13, or 14 under stringent conditions; and,

c) a fragment of a polypeptide comprising the amino acid sequence of SEQ ID NO:1, 3, 5, or 7, or the amino acid sequence encoded by the cDNA insert of the plasmid deposited with the ATCC as Patent Deposit Number ____, PTA-1639, PTA-1846, or ____, wherein the fragment comprises at least 15 contiguous amino acids of SEQ ID NO:1, 3, 5, or 7; and

d) a polypeptide having at least 60% sequence identity to the amino acid sequence SEQ ID NO:1, 3, 5, or 7, wherein the polypeptide has biological activity.

9. The isolated polypeptide of claim 8 comprising the amino acid sequence of SEQ ID NO:1, 3, 5, or 7, or an amino acid sequence encoded by the cDNA insert of any of the plasmids deposited with ATCC as Patent Deposit Number ____, PTA-1639, PTA-1846, or ____.

10. The polypeptide of claim 8 further comprising heterologous amino acid sequences.

11. An antibody which selectively binds to a polypeptide of claim 8.

12. A method for producing a polypeptide selected from the group consisting of:

a) a polypeptide comprising the amino acid sequence of SEQ ID NO:1, 3, 5, or 7, or the amino acid sequence encoded by the cDNA insert of the

plasmid deposited with the ATCC as Patent Deposit Number ____, PTA-1639, PTA-1846, or ____;

b) a polypeptide comprising a fragment of the amino acid sequence of SEQ ID NO:1, 3, 5, or 7, or the amino acid sequence encoded by the cDNA insert of the plasmid deposited with the ATCC as Patent Deposit Number ____, PTA-1639, PTA-1846, or ____, wherein the fragment comprises at least 15 contiguous amino acids of SEQ ID NO:1, 3, 5, or 7, or the amino acid sequence encoded by the cDNA insert of the plasmid deposited with the ATCC as Patent Deposit Number ____, PTA-1639, PTA-1846, or ____;

c) a biologically active naturally occurring allelic variant of a polypeptide comprising the amino acid sequence of SEQ ID NO:1, 3, 5, or 7, or the amino acid sequence encoded by the cDNA insert of the plasmid deposited with the ATCC as Patent Deposit Number ____, PTA-1639, PTA-1846, or ____, wherein the polypeptide is encoded by a nucleic acid molecule which hybridizes to a nucleic acid molecule comprising the complement of SEQ ID NO:2, 4, 6, 8, 11, 12, 13, or 14;

d) a polypeptide having at least 60% sequence identity to the amino acid sequence of SEQ ID NO:1, 3, 5, or 7, wherein said polypeptide has biological activity;

comprising culturing the host cell of claim 5 under conditions in which the nucleic acid molecule is expressed.

13. The method of claim 12 wherein said polypeptide comprises the amino acid sequence of SEQ ID NO:1, 3, 5, or 7.

14. A method for detecting the presence of a polypeptide of claim 8 in a sample, comprising:

a) contacting the sample with a compound which selectively binds to a polypeptide of claim 8; and

b) determining whether the compound binds to the polypeptide in the sample.

15. The method of claim 14, wherein the compound which binds to the polypeptide is an antibody.

16. A kit comprising a compound which selectively binds to a polypeptide of claim 8 and instructions for use.

17. A method for detecting the presence of a nucleic acid molecule of claim 1 in a sample, comprising the steps of:

a) contacting the sample with a nucleic acid probe or primer which selectively hybridizes to the nucleic acid molecule; and

b) determining whether the nucleic acid probe or primer binds to a nucleic acid molecule in the sample.

18. The method of claim 17, wherein the sample comprises mRNA molecules and is contacted with a nucleic acid probe.

19. A kit comprising a compound which selectively hybridizes to a nucleic acid molecule of claim 1 and instructions for use.

20. A method for identifying a compound which binds to a polypeptide of claim 8 comprising the steps of:

a) contacting a polypeptide, or a cell expressing a polypeptide of claim 8 with a test compound; and

b) determining whether the polypeptide binds to the test compound.

21. The method of claim 20, wherein the binding of the test compound to the polypeptide is detected by a method selected from the group consisting of:

a) detection of binding by direct detecting of test compound/polypeptide binding;

b) detection of binding using a competition binding assay;

c) detection of binding using an assay for sulfatase activity.

22. A method for modulating the activity of a polypeptide of claim 8 comprising contacting a polypeptide or a cell expressing a polypeptide of claim 8 with a compound which binds to the polypeptide in a sufficient concentration to modulate the activity of the polypeptide.

23. A method for identifying a compound which modulates the activity of a polypeptide of claim 8, comprising:

a) contacting a polypeptide of claim 8 with a test compound; and

b) determining the effect of the test compound on the activity of the polypeptide to thereby identify a compound which modulates the activity of the polypeptide.

24. A method for identifying an agent that modulates the level of expression of a nucleic acid molecule of claim 1 in a cell, said method comprising contacting said agent with the cell expressing said nucleic acid molecule such that said level of expression of said nucleic acid molecule can be modulated in said cell by said agent and measuring said level of expression of said nucleic acid molecule.

25. A method for modulating the level of expression of a nucleic acid molecule of claim 1, said method comprising contacting said nucleic acid molecule with an agent under conditions that allow the agent to modulate the level of expression of the nucleic acid molecule.

26. A pharmaceutical composition containing any of the polypeptides in claim 8 in a pharmaceutically acceptable carrier.

Sequence length 2175

CACGCGTCCGCAAATTTCCTGATTCTTTTGAATTAGGATTCCAGATGGGGGCCTCATTTCTACAGCCCCCAACATTCCT

ATAGCCGTTATCACTGCCATCACCACTGCCACCAGCATCTTCTTGCAGATTCCACCCCTGCTCCCCAGAGACTTCCTGC

TTTGAAAGTGAGCAGAAAGGAAGCTCTCAGAAAAATCTCTAGTGGTGGCTGCCGTCGCTCCAGACAATCGGAATCCTGC

```
                M   G   W   L   F   L   K   V   L   L   A   G   V   S   F   S   G    17
CTTCACCACC ATG GGC TGG CTT TTT CTA AAG GTT TTG TTG GCG GGA GTG AGT TTC TCA GGA      51

 F   L   Y   P   L   V   D   F   C   I   S   G   K   T   R   G   Q   K   P   N    37
TTT CTT TAT CCT CTT GTG GAT TTT TGC ATC AGT GGG AAA ACA AGA GGA CAG AAG CCA AAC    111

 F   V   I   I   L   A   D   D   M   G   W   G   D   L   G   A   N   W   A   E    57
TTT GTG ATT ATT TTG GCC GAT GAC ATG GGG TGG GGT GAC CTG GGA GCA AAC TGG GCA GAA    171

 T   K   D   T   A   N   L   D   K   M   A   S   E   G   M   R   F   V   D   F    77
ACA AAG GAC ACT GCC AAC CTT GAT AAG ATG GCT TCG GAG GGA ATG AGG TTT GTG GAT TTC    231

 H   A   A   A   S   T   C   S   P   S   R   A   S   L   L   T   G   R   L   G    97
CAT GCA GCT GCC TCC ACC TGC TCA CCC TCC CGG GCT TCC TTG CTC ACC GGC CGG CTT GGC    291

 L   R   N   G   V   T   R   N   F   A   V   T   S   V   G   G   L   P   L   N   117
CTT CGC AAT GGA GTC ACA CGC AAC TTT GCA GTC ACT TCT GTG GGA GGC CTT CCG CTC AAC    351

 E   T   T   L   A   E   V   L   Q   Q   A   G   Y   V   T   G   I   I   G   K   137
GAG ACC ACC TTG GCA GAG GTG CTG CAG CAG GCG GGT TAC GTC ACT GGG ATA ATA GGC AAA    411

 W   H   L   G   H   H   G   S   Y   H   P   N   F   R   G   F   D   Y   Y   F   157
TGG CAT CTT GGA CAC CAC GGC TCT TAT CAC CCC AAC TTC CGT GGT TTT GAT TAC TAC TTT    471

 G   I   P   Y   S   H   D   M   G   C   T   D   T   P   G   Y   N   H   P   P   177
GGA ATC CCA TAT AGC CAT GAT ATG GGC TGT ACT GAT ACT CCA GGC TAC AAC CAC CCT CCT    531

 C   P   A   C   P   Q   G   D   G   P   S   R   N   L   Q   R   D   C   Y   T   197
TGT CCA GCG TGT CCA CAG GGT GAT GGA CCA TCA AGG AAC CTT CAA AGA GAC TGT TAC ACT    591

 D   V   A   L   P   L   Y   E   N   L   N   I   V   E   Q   P   V   N   L   S   217
GAC GTG GCC CTC CCT CTT TAT GAA AAC CTC AAC ATT GTG GAG CAG CCG GTG AAC TTG AGC    651

 S   L   A   Q   K   Y   A   E   K   A   T   Q   F   I   Q   R   A   S   T   S   237
AGC CTT GCC CAG AAG TAT GCT GAG AAA GCA ACC CAG TTC ATC CAG CGT GCA AGC ACC AGC    711

 G   R   P   F   L   L   Y   V   A   L   A   H   M   H   V   P   L   P   V   T   257
GGG AGG CCC TTC CTG CTC TAT GTG GCT CTG GCC CAC ATG CAC GTG CCC TTA CCC GTG ACT    771

 Q   L   P   A   A   P   R   G   R   S   L   Y   G   A   G   L   W   E   M   D   277
CAG CTA CCA GCA GCG CCA CGG GGC AGA AGC CTG TAT GGT GCA GGG CTC TGG GAG ATG GAC    831

 S   L   V   G   Q   I   K   D   K   V   D   H   T   V   K   E   N   T   F   L   297
AGT CTG GTG GGC CAG ATC AAG GAC AAA GTT GAC CAC ACA GTG AAG GAA AAC ACA TTC CTC    891

 W   F   T   G   D   N   G   P   W   A   Q   K   C   E   L   A   G   S   V   G   317
TGG TTT ACA GGA GAC AAT GGC CCG TGG GCT CAG AAG TGT GAG CTA GCG GGC AGT GTG GGT    951

 P   F   T   G   F   W   Q   T   R   Q   G   G   S   P   A   K   Q   T   T   W   337
CCC TTC ACT GGA TTT TGG CAA ACT CGT CAA GGG GGA AGT CCA GCC AAG CAG ACG ACC TGG   1011

 E   G   G   H   R   V   P   A   L   A   Y   W   P   G   R   V   P   V   N   V   357
```

FIGURE 1A

```
GAA GGA GGG CAC CGG GTC CCA GCA CTG GCT TAC TGG CCT GGC AGA GTT CCA GTT AAT GTC  1071

 T   S   T   A   L   L   S   V   L   D   I   F   P   T   V   V   A   L   A   Q   377
ACC AGC ACT GCC TTG TTA AGC GTG CTG GAC ATT TTT CCA ACT GTG GTA GCC CTG GCC CAG  1131

 A   S   L   P   Q   G   R   R   F   D   G   V   D   V   S   E   V   L   F   G   397
GCC AGC TTA CCT CAA GGA CGG CGC TTT GAT GGT GTG GAC GTC TCC GAG GTG CTC TTT GGC  1191

 R   S   Q   P   G   H   R   V   L   F   H   P   N   S   G   A   A   G   E   F   417
CGG TCA CAG CCT GGG CAC AGG GTG CTG TTC CAC CCC AAC AGC GGG GCA GCT GGA GAG TTT  1251

 G   A   L   Q   T   V   R   L   E   R   Y   K   A   F   Y   I   T   G   G   A   437
GGA GCC CTG CAG ACT GTC CGC CTG GAG CGT TAC AAG GCC TTC TAC ATT ACC GGT GGA GCC  1311

 R   A   C   D   G   S   T   G   P   E   L   Q   H   K   F   P   L   I   F   N   457
AGG GCG TGT GAT GGG AGC ACG GGG CCT GAG CTG CAG CAT AAG TTT CCT CTG ATT TTC AAC  1371

 L   E   D   D   T   A   E   A   V   P   L   E   R   G   G   A   E   Y   Q   A   477
CTG GAA GAC GAT ACC GCA GAA GCT GTG CCC CTA GAA AGA GGT GGT GCG GAG TAC CAG GCT  1431

 V   L   P   E   V   R   K   V   L   A   D   V   L   Q   D   I   A   N   D   N   497
GTG CTG CCC GAG GTC AGA AAG GTT CTT GCA GAC GTC CTC CAA GAC ATT GCC AAC GAC AAC  1491

 I   S   S   A   D   Y   T   Q   D   P   S   V   T   P   C   C   N   P   Y   Q   517
ATC TCC AGC GCA GAT TAC ACT CAG GAC CCT TCA GTA ACT CCC TGC TGT AAT CCC TAC CAA  1551

 I   A   C   R   C   Q   A   A   *                                               526
ATT GCC TGC CGC TGT CAA GCC GCA TAA                                              1578

CAGACCAATTTTTATTCCACGAGGAGGAGTACCTGGAAATTAGGCAAGTTTGCTTCCAAATTTCATTTTTACCCTCTTT

ACAAACACACGCTTTAGTTTAGTCTTGGAGTTTAGTTTTGGAGTTAGCCTTGCATATCCCTTCTGTATCCTGTCCCTCC

TCCACGCCGACCCGAGAGCAGCTGAGCTGCGCTGGCTCTGGGCAGGGAGTGTGCCTTAATGGGAAGCACACGGGCTTTG

GAGTCAGGCACAGGTGCCAGCTCCAGCTTTTGAACTTGGGCAATTGTTTAACCTAACCTGCAAGTTGATTTTGAGGGTT

AAATAAAGGCATACATGAAAAAAAAAAAAAAAAA
```

FIGURE 1B

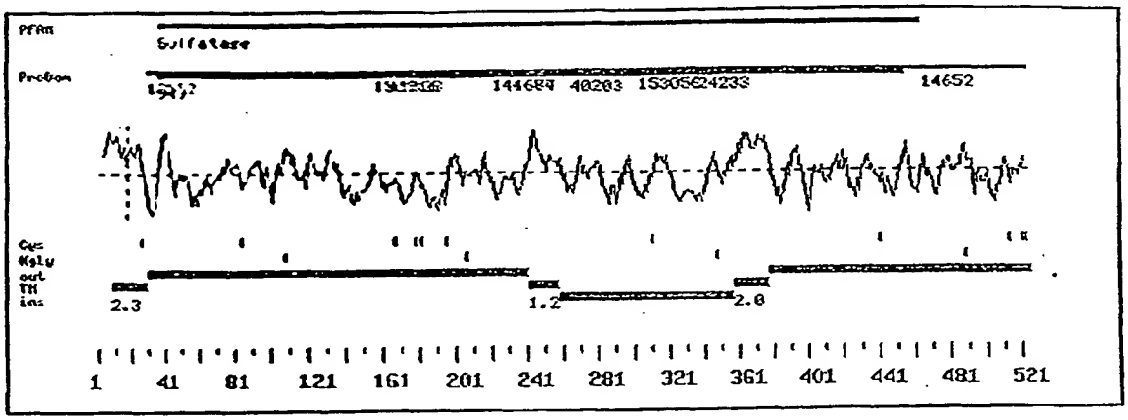
PFAM
Sulfatase
Prodom
144684
40283
14652
2.3
1.2
2.0
1
41
81
121
161
201
241
281
321
361
401
441
481
521

FIGURE 2

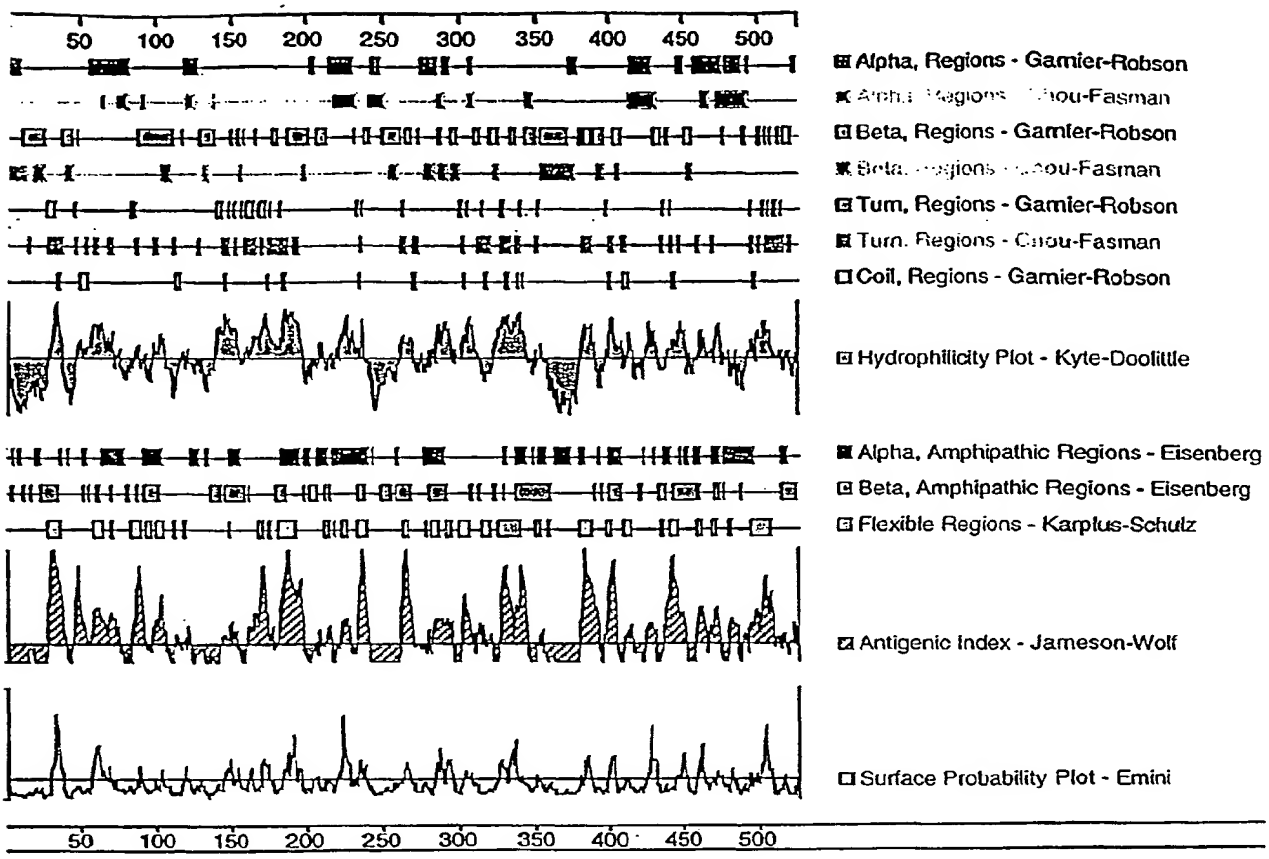

FIGURE 3

Prosite Pattern Matches

Prosite version: Release 12.2 of February 1995

>PS00001|PDOC00001|ASN_GLYCOSYLATION N-glycosylation site.

Query: 117 NETT 120
Query: 215 NLSS 218
Query: 356 NVTS 359
Query: 497 NISS 500

>PS00005|PDOC00005|PKC_PHOSPHO_SITE Protein kinase C phosphorylation site.

Query: 28 SGK 30
Query: 93 TGR 95
Query: 237 SGR 239
Query: 290 TVK 292
Query: 422 TVR 424

>PS00006|PDOC00006|CK2_PHOSPHO_SITE Casein kinase II phosphorylation site.

Query: 120 TLAE 123
Query: 290 TVKE 293
Query: 335 TTWE 338
Query: 364 SVLD 367
Query: 444 TGPE 447
Query: 499 SSAD 502

>PS00008|PDOC00008|MYRISTYL N-myristoylation site.

Query: 12 GVSFSG 17
Query: 33 GQKPNF 38
Query: 52 GANWAE 57
Query: 97 GLRNGV 102
Query: 113 GLPLNE 118
Query: 158 GIPYSH 163
Query: 328 GGSPAK 333
Query: 388 GVDVSE 393
Query: 418 GALQTV 423
Query: 435 GGARAC 440

>PS00009|PDOC00009|AMIDATION Amidation site.

Query: 382 QGRR 385

>PS00149|PDOC00117|SULFATASE_2 Sulfatases signature 2.

Query: 129 GYVTGIIGKW 138

FIGURE 4

Input file Fbh23553fl.seq; Output File 23553.trans
Sequence length 4321

CCCACGCGTCCGGCTAATGAATCTTGGGGCCGGTGTCGGGCCGGGGCGGCTTGATCGGCAACTAGGAAACCCCAGGCGC

AGAGGCCAGGAGCGAGGGCAGCGAGGATCAGAGGCCAGGCCTTCCCGGCTGCCGGCGCTCCTCGGAGGTCAGGGCAGAT

GAGGAACATGACTCTCCCCCTTCGGAGGAGGAAGGAAGTCCCGCTGCCACCTTATCTCTGCTCCTCTGCCTCCTCCCTG

TTCCCAGAGCTTTTTCTCTAGAGAAGATTTTGAAGGCGGCTTTTGTGCTGACGGCCACCCACCATCATCTAAAGAAGAT

AAACTTGGCAAATGACATGCAGGTTCTTCAAGGCAGAATAATTGCAGAAAATCTTCAAAGGACCCTATCTGCAGATGTT

CTCAATACCTCTGAGAATAGAGATTGATTATTCAACCAGGATACCTAATTCAAGAACTCCAGAAATCAGGAGACGGAGA

	M K Y S C C A L V L A	11
CATTTTGTCAGTTTTGCAACATTGGACCAAATACA ATG AAG TAT TCT TGC TGT GCT CTG GTT TTG GCT		33
V L G T E L L G S L C S T V R S P R F R		31
CTC CTC GGC ACA GAA TTG CTG GGA AGC CTC TGT TCG ACT GTC AGA TCC CCG AGG TTC AGA		93
G R I Q Q E R K N I R P N I I L V L T D		51
GGA CGG ATA CAG CAG GAA CGA AAA AAC ATC CGA CCC AAC ATT ATT CTT GTG CTT ACC GAT		153
D C D V E L G S L Q V M N K T R K I M E		71
GAT CAA GAT GTC GAG CTG GGG TCC CTG CAA GTC ATG AAC AAA ACG AGA AAG ATT ATG GAA		213
E G G A T F I N A F V T T P M C C P S R		91
GAA GGG GGG GCC ACC TTC ATC AAT GCC TTT GTG ACT ACA CCC ATG TGC TGC CCG TCA CGG		273
S S M L T G K Y V H N H N V Y T N N E N		111
TCC TCC ATG CTC ACC GGG AAG TAT GTG CAC AAT CAC AAT GTC TAC ACC AAC AAC GAG AAC		333
C S S P S W Q A M H E P R T F A V Y L N		131
TGC TCT TCC CCC TCG TGG CAG GCC ATG CAT GAG CCT CGG ACT TTT GCT GTA TAT CTT AAC		393
N T G Y R T A F F G K Y L N E Y N G S Y		151
AAC ACT GGC TAC AGA ACA GCC TTT TTT GGA AAA TAC CTC AAT GAA TAT AAT GGC AGC TAC		453
I P P G W R E W L G L I K N S R F Y N Y		171
ATC CCC CCT GGG TGG CGA GAA TGG CTT GGA TTA ATC AAG AAT TCT CGC TTC TAT AAT TAC		513
T V C R N G I K E K H G F D Y A K D Y F		191
ACT GTT TGT CGC AAT GGC ATC AAA GAA AAG CAT GGA TTT GAT TAT GCA AAG GAC TAC TTC		573
T D L I T N E S I N Y F K M S K R M Y P		211
ACA GAC TTA ATC ACT AAC GAG AGC ATT AAT TAC TTC AAA ATG TCT AAG AGA ATG TAT CCC		633
H R P V M M V I S H A A P H G P E D S A		231
CAT AGG CCC GTT ATG ATG GTG ATC AGC CAC GCT GCG CCC CAC GGC CCC GAG GAC TCA GCC		693
P Q F S K L Y P N A S Q H I T P S Y N Y		251
CCA CAG TTT TCT AAA CTG TAC CCC AAT GCT TCC CAA CAC ATA ACT CCT AGT TAT AAC TAT		753
A P N M D K H W I M Q Y T G P M L P I H		271
GCA CCA AAT ATG GAT AAA CAC TGG ATT ATG CAG TAC ACA GGA CCA ATG CTG CCC ATC CAC		813
M E F T N I L Q R K R L Q T L M S V D D		291
ATG GAA TTT ACA AAC ATT CTA CAG CGC AAA AGG CTC CAG ACT TTG ATG TCA GTG GAT GAT		873
S V E R L Y N M L V E T G E L E N T Y I		311

FIGURE 5A

```
TCT GTG GAG AGG CTG TAT AAC ATG CTC GTG GAG ACG GGG GAG CTG GAG AAT ACT TAC ATC   933

 I  Y  T  A  D  H  G  Y  H  I  G  Q  F  G  L  V  K  G  K  S                      331
ATT TAC ACC GCC GAC CAT GGT TAC CAT ATT GGG CAG TTT GGA CTG GTC AAG GGG AAA TCC   993

 M  P  Y  D  F  D  I  R  V  P  F  F  I  R  G  P  S  V  E  P                     351
ATG CCA TAT GAC TTT GAT ATT CGT GTG CCT TTT TTT ATT CGT GGT CCA AGT GTA GAA CCA  1053

 G  S  I  V  P  Q  I  V  L  N  I  D  L  A  P  T  I  L  D  I                     371
GGA TCA ATA GTC CCA CAG ATC GTT CTC AAC ATT GAC TTG GCC CCC ACG ATC CTG GAT ATT  1113

 A  G  L  D  T  P  P  D  V  D  G  K  S  V  L  K  L  L  D  P                     391
GCT GGG CTC GAC ACA CCT CCT GAT GTG GAC GGC AAG TCT GTC CTC AAA CTT CTG GAC CCA  1173

 E  K  P  G  N  R  F  R  T  N  K  K  A  K  I  W  R  D  T  F                     411
GAA AAG CCA GGT AAC AGG TTT CGA ACA AAC AAG AAG GCC AAA ATT TGG CGT GAT ACA TTC  1233

 L  V  E  R  G  K  F  L  R  K  K  E  E  S  S  K  N  I  Q  Q                     431
CTA GTG GAA AGA GGC AAA TTT CTA CGT AAG AAG GAA GAA TCC AGC AAG AAT ATC CAA CAG  1293

 S  N  H  L  P  K  Y  E  R  V  K  E  L  C  Q  Q  A  R  Y  Q                     451
TCA AAT CAC TTG CCC AAA TAT GAA CGG GTC AAA GAA CTA TGC CAG CAG GCC AGG TAC CAG  1353

 T  A  C  E  Q  P  G  Q  K  W  Q  C  I  E  D  T  S  G  K  L                     471
ACA GCC TGT GAA CAA CCG GGG CAG AAG TGG CAA TGC ATT GAG GAT ACA TCT GGC AAG CTT  1413

 R  I  H  K  C  K  G  P  S  D  L  L  T  V  R  Q  S  T  R  N                     491
CGA ATT CAC AAG TGT AAA GGA CCC AGT GAC CTG CTC ACA GTC CGG CAG AGC ACG CGG AAC  1473

 L  Y  A  R  G  F  H  D  K  D  K  E  C  S  C  R  E  S  G  Y                     511
CTC TAC GCT CGC GGC TTC CAT GAC AAA GAC AAA GAG TGC AGT TGT AGG GAG TCT GGT TAC  1533

 R  A  S  R  S  Q  R  K  S  Q  R  Q  F  L  R  N  Q  G  T  P                     531
CGT GCC AGC AGA AGC CAA AGA AAG AGT CAA CGG CAA TTC TTG AGA AAC CAG GGG ACT CCA  1593

 K  Y  K  P  R  F  V  H  T  R  Q  T  R  S  L  S  V  E  F  E                     551
AAG TAC AAG CCC AGA TTT GTC CAT ACT CGG CAG ACA CGT TCC TTG TCC GTC GAA TTT GAA  1653

 G  E  I  Y  D  I  N  L  E  E  E  E  E  L  Q  V  L  Q  P  R                     571
GGT GAA ATA TAT GAC ATA AAT CTG GAA GAA GAA GAA GAA TTG CAA GTG TTG CAA CCA AGA  1713

 N  I  A  K  R  H  D  E  G  H  K  G  P  R  D  L  Q  A  S  S                     591
AAC ATT GCT AAG CGT CAT GAT GAA GGC CAC AAG GGG CCA AGA GAT CTC CAG GCT TCC AGT  1773

 G  G  N  R  G  R  M  L  A  D  S  S  N  A  V  G  P  P  T  T                     611
GGT GGC AAC AGG GGC AGG ATG CTG GCA GAT AGC AGC AAC GCC GTG GGC CCA CCT ACC ACT  1833

 V  R  V  T  H  K  C  F  I  L  P  N  D  S  I  H  C  E  R  E                     631
GTC CGA GTG ACA CAC AAG TGT TTT ATT CTT CCC AAT GAC TCT ATC CAT TGT GAG AGA GAA  1893

 L  Y  Q  S  A  R  A  W  K  D  H  K  A  Y  I  D  K  E  I  E                     651
CTG TAC CAA TCG GCC AGA GCG TGG AAG GAC CAT AAG GCA TAC ATT GAC AAA GAG ATT GAA  1953

 A  L  Q  D  K  I  K  N  L  R  E  V  R  G  H  L  K  R  R  K                     671
GCT CTG CAA GAT AAA ATT AAG AAT TTA AGA GAA GTG AGA GGA CAT CTG AAG AGA AGG AAG  2013

 P  E  E  C  S  C  S  K  Q  S  Y  Y  N  K  E  K  G  V  K  K                     691
CCT GAG GAA TGT AGC TGC AGT AAA CAA AGC TAT TAC AAT AAA GAG AAA GGT GTA AAA AAG  2073

 Q  E  K  L  K  S  H  L  H  P  F  K  E  A  A  Q  E  V  D  S                     711
CAA GAG AAA TTA AAG AGC CAT CTT CAC CCA TTC AAG GAG GCT GCT CAG GAA GTA GAT AGC  2133
```

FIGURE 5B

```
 K  L  Q  L  F  K  E  N  N  R  R  R  K  K  E  R  K  E  K  R      731
AAA CTG CAA CTT TTC AAG GAG AAC AAC CGT AGG AGG AAG AAG GAG AGG AAG GAG AAG AGA    2193

 R  Q  R  K  G  E  E  C  S  L  P  G  L  T  C  F  T  H  D  N      751
CGG CAG AGG AAG GGG GAA GAG TGC AGC CTG CCT GGC CTC ACT TGC TTC ACG CAT GAC AAC    2253

 N  H  W  Q  T  A  P  F  W  N  L  G  S  F  C  A  C  T  S  S      771
AAC CAC TGG CAG ACA GCC CCG TTC TGG AAC CTG GGA TCT TTC TGT GCT TGC ACG AGT TCT    2313

 N  N  N  T  Y  W  C  L  R  T  V  N  E  T  H  N  F  L  F  C      791
AAC AAT AAC ACC TAC TGG TGT TTG CGT ACA GTT AAT GAG ACG CAT AAT TTT CTT TTC TGT    2373

 E  F  A  T  G  F  L  E  Y  F  D  M  N  T  D  P  Y  Q  L  T      811
GAG TTT GCT ACT GGC TTT TTG GAG TAT TTT GAT ATG AAT ACA GAT CCT TAT CAG CTC ACA    2433

 N  T  V  H  T  V  E  R  G  I  L  N  Q  L  H  V  Q  L  M  E      831
AAT ACA GTG CAC ACG GTA GAA CGA GGC ATT TTG AAT CAG CTA CAC GTA CAA CTA ATG GAG    2493

 L  R  S  C  Q  G  Y  K  Q  C  N  P  R  P  K  N  L  D  V  G      851
CTC AGA AGC TGT CAA GGA TAT AAG CAG TGC AAC CCA AGA CCT AAG AAT CTT GAT GTT GGA    2553

 N  K  D  G  G  S  Y  D  L  H  R  G  Q  L  W  D  G  W  E  G      871
AAT AAA GAT GGA GGA AGC TAT GAC CTA CAC AGA GGA CAG TTA TGG GAT GGA TGG GAA GGT    2613

 *
TAA                                                                872
                                                                   2616

TCAGCCCCGTCTCACTGCAGACATCAACTGGCAAGGCCTAGAGGAGCTACACAGTGTGAATGAAAACATCTATGAGTAC

AGACAAAACTACAGACTTAGTCTGGTGGACTGGACTAATTACTTGAAGGATTTAGATAGAGTATTTGCACTGCTGAAGA

GTCACTATGAGCAAAATAAAACAAATAAGACTCAAACTGCTCAAAGTGACGGGTTCTTGGTTGTCTCTGCTGAGCACGC

TGTGTCAATGGAGATGGCCTCTGCTGACTCAGATGAAGACCCAAGGCATAAGGTTGGGAAAACACCTCATTTGACCTTG

CCAGCTGACCTTCAAACCCTGCATTTGAACCGACCAACATTAAGTCCAGAGAGTAAACTTGAATGGAATAACGACATTC

CAGAAGTTAATCATTTGAATTCTGAACACTGGAGAAAAACCGAAAAATGGACGGGGCATGAAGAGACTAATCATCTGGA

AACCGATTTCAGTGGCGATGGCATGACAGAGCTAGAGCTCGGGCCCAGCCCCAGGCTGCAGCCCATTCGCAGGCACCCG

AAAGAACTTCCCCAGTATGGTGGTCCTGGAAAGGACATTTTTGAAGATCAACTATATCTTCCTGTGCATTCCGATGGAA

TTTCAGTTCATCAGATGTTCACCATGGCCACCGCAGAACACCGAAGTAATTCCAGCATAGCGGGGAAGATGTTGACCAA

GGTGGAGAAGAATCACGAAAAGGAGAAGTCACAGCACCTAGAAGGCAGCGCCTCCTCTTCACTCTCCTCTGATTAGATG

AAACTGTTACCTTACCCTAAACACAGTATTTCTTTTTAACTTTTTTATTTGTAAACTAATAAAGGKAATCACWGCCACC

AACATTCCAAGCTACCCTGGGTACCTTTGTGCAGTAGAAGCTAGTGAGCATGTGAGCAAGCGGTGTGCACACGGAGACT

CATCGTTATAATTTACTATCTGCCAAGGAGTAGAAAGAAAGGCTGGGGATATTTGGGTTGGCTTTGGKTTTGATTTTTT

GCTTGGTTGGTTGGTTTGKACTAAAACAGTATTATCTTTTGAATATCGTAGGGACATAARKWWWWMMWKKIWWTCMAW

YMRAKAKGSYWRRAWKGGGSTYTYTSKKRKSTMWAMWYKWSCHCCYSKKRWWAWTYWYWMMYWCMYKYTSSSTGRYKRN

KTAATGAAGTT
```

FIGURE 5C

Analysis of 23553 (871 aa)

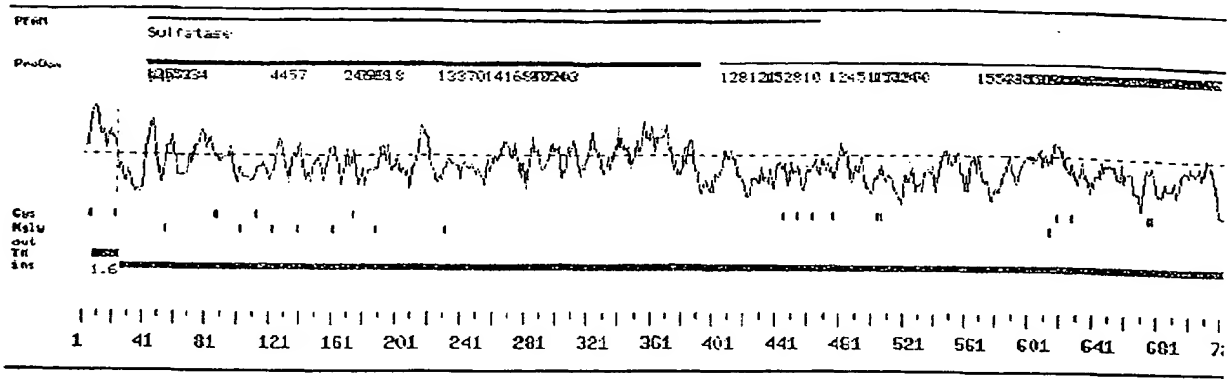

FIGURE 6

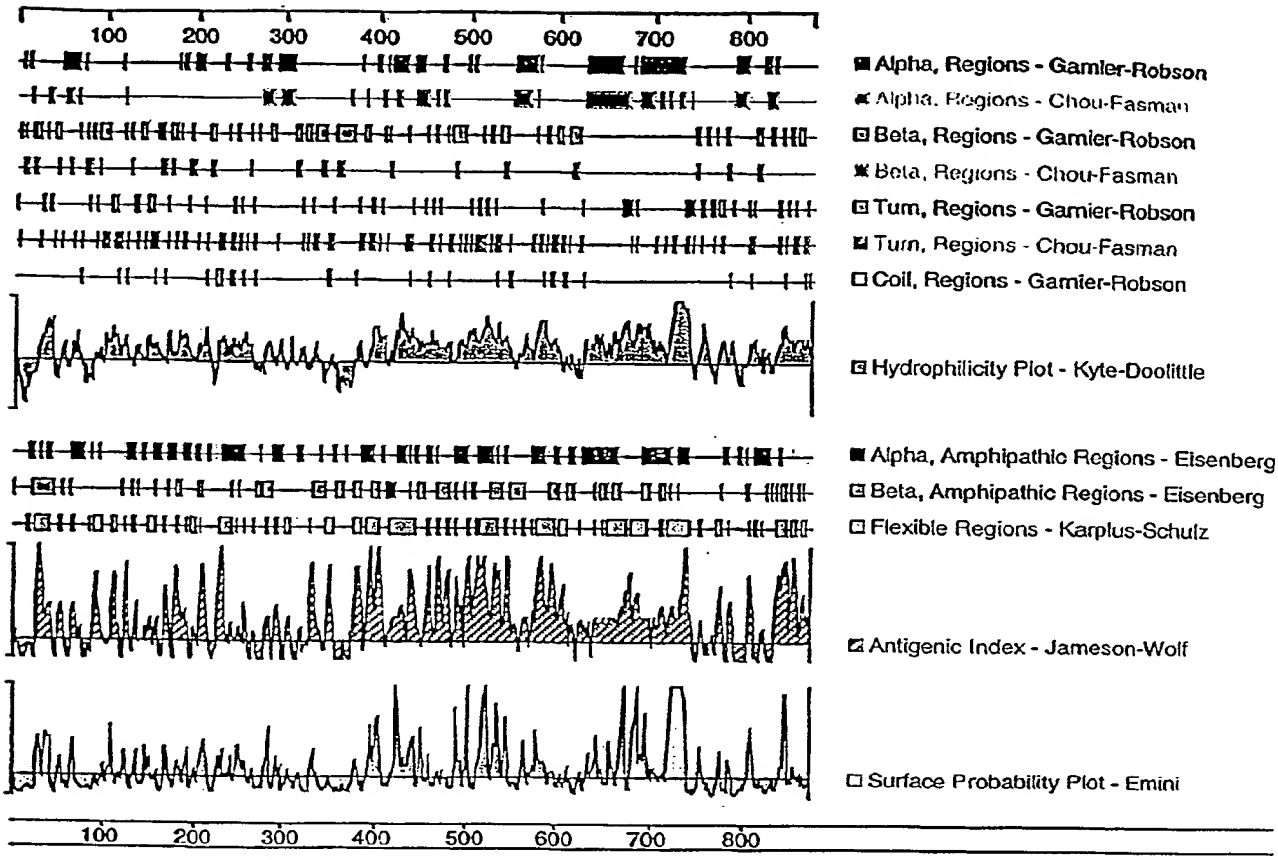
100
200
300
400
500
600
700
800
Alpha, Regions - Gamier-Robson
Alpha, Regions - Chou-Fasman
Beta, Regions - Gamier-Robson
Beta, Regions - Chou-Fasman
Turn, Regions - Gamier-Robson
Turn, Regions - Chou-Fasman
Coil, Regions - Gamier-Robson
Hydrophilicity Plot - Kyte-Doolittle
Alpha, Amphipathic Regions - Eisenberg
Beta, Amphipathic Regions - Eisenberg
Flexible Regions - Karplus-Schulz
Antigenic Index - Jameson-Wolf
Surface Probability Plot - Emini
100
200
300
400
500
600
700
800

FIGURE 7

Prosite Pattern Matches for 23553

Prosite version: Release 12.2 of February 1995

>PS00001|PDOC00001|ASN_GLYCOSYLATION N-glycosylation site.

Query:	64	NKTR	67
Query:	111	NCSS	114
Query:	131	NNTG	134
Query:	148	NGSY	151
Query:	170	NYTV	173
Query:	197	NESI	200
Query:	240	NASQ	243
Query:	623	NDSI	626
Query:	773	NNTY	776
Query:	783	NETH	786

>PS00005|PDOC00005|PKC_PHOSPHO_SITE Protein kinase C phosphorylation site.

Query:	24	TVR	26
Query:	27	SPR	29
Query:	66	TRK	68
Query:	96	TGK	98
Query:	206	SKR	208
Query:	400	TNK	402
Query:	425	SSK	427
Query:	468	SGK	470
Query:	484	TVR	486
Query:	488	STR	490
Query:	505	SCR	507
Query:	516	SQR	518
Query:	520	SQR	522
Query:	530	TPK	532
Query:	611	TVR	613
Query:	615	THK	617
Query:	635	SAR	637

>PS00006|PDOC00006|CK2_PHOSPHO_SITE Casein kinase II phosphorylation site.

Query:	107	TNNE	110
Query:	288	SVDD	291
Query:	367	TILD	370
Query:	376	TPPD	379
Query:	452	TACE	455
Query:	505	SCRE	508
Query:	781	TVNE	784

FIGURE 8A

>PS00007|PDOC00007|TYR_PHOSPHO_SITE Tyrosine kinase phosphorylation site.

Query: 637 RAWKDHKAY 645

>PS00008|PDOC00008|MYRISTYL N-myristoylation site.

Query: 19 GSLCST 24

Query: 161 GLIKNS 166

Query: 325 GLVKGK 330

Query: 592 GGNRGR 597

Query: 763 GSFCAC 768

Query: 851 GNKDGG 856

>PS00523|PDOC00117|SULFATASE_1 Sulfatases signature 1.

Query: 85 PHCCPSRSSMLTG 97

FIGURE 8B

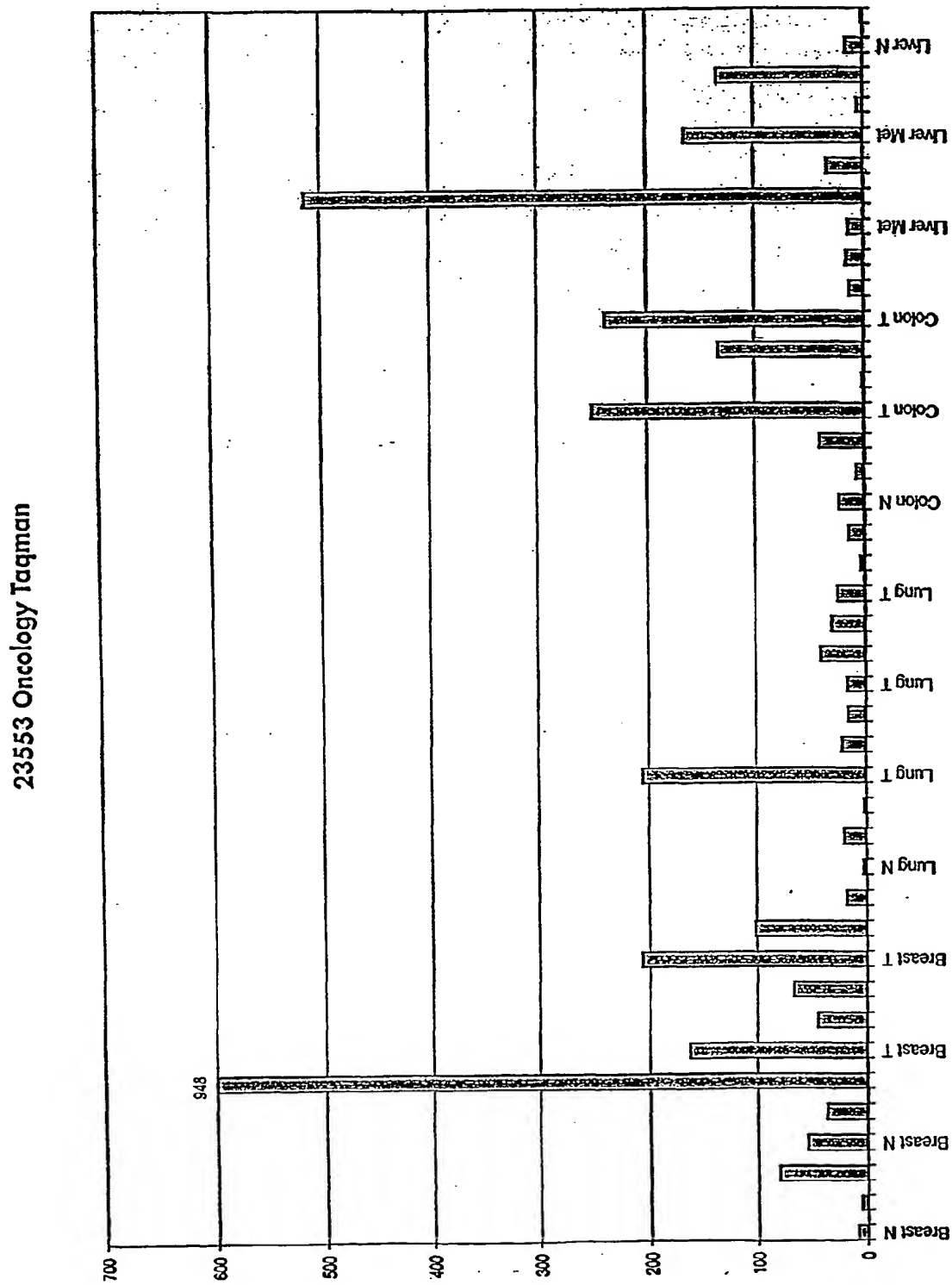

FIGURE 9

```
Input file Fbh25278FL1.seq; Output File 25278.trans
Sequence length 2940

CCACGCGTCCGCCCACGCGTCCGGCTGCCACGCCGCGTCTCAGGCTGGCCGGGCTGAGCCGGGGAAGAGGGAGCAAAGG

CGGCGCAGGGCCTGCGCTTAGGCAGCGGGAGGCAGCTCGGCGCGGGCCTGACCTCCCCAGAGCGCCCCGCTGCGGCCGA

GCAGATCCGGCCCAGCCGTCCGGCAGCCAGTCCCGGACCAGACACTGGACCGTCCCCGGGGGGCGCTGAACTCCCTCGC

AGCATCCGAGCCGGCGGGCCGGTGGTGCGCCCTGGGCGCGCGAGGTGGTGAGGCCCCAGGAGCCCGGCGCGCCGGGACA

                    M  H  T  L  T  G  F  S  L  V  S  L  L  S  F        15
CGCGGGCCGGCTTGGCG ATG CAC ACC CTC ACT GGC TTC TCT CTG GTC AGC CTG CTC AGC TTC      45

 G  Y  L  S  W  D  W  A  K  P  S  F  V  A  D  G  P  G  E  A        35
GGC TAC CTG TCC TGG GAC TGG GCC AAG CCG AGC TTC GTG GCC GAC GGG CCC GGG GAG GCT   105

 G  E  Q  P  S  A  A  P  P  Q  P  P  H  I  I  F  I  L  T  D        55
GGC GAG CAG CCC TCG GCC GCT CCG CCC CAG CCT CCC CAC ATC ATC TTC ATC CTC ACG GAC   165

 D  Q  G  Y  H  D  V  G  Y  H  G  S  D  I  E  T  P  T  L  D        75
GAC CAA GGC TAC CAC GAC GTG GGC TAC CAT GGT TCA GAT ATC GAG ACC CCT ACG CTG GAC   225

 R  L  A  A  K  G  V  K  L  E  N  Y  Y  I  Q  P  I  C  T  P        95
AGG CTG GCG GCC AAG GGG GTC AAG TTG GAG AAT TAT TAC ATC CAG CCC ATC TGC ACG CCT   285

 S  R  S  Q  L  L  T  G  R  Y  Q  I  H  T  G  L  Q  H  S  I       115
TCG CGG AGC CAG CTC CTC ACT GGC AGG TAC CAG ATC CAC ACA GGA CTC CAG CAT TCC ATC   345

 I  R  P  Q  Q  P  N  C  L  P  L  D  Q  V  T  L  P  Q  K  L       135
ATC CGC CCA CAG CAG CCC AAC TGC CTG CCC CTG GAC CAG GTG ACA CTG CCA CAG AAG CTG   405

 Q  E  A  G  Y  S  T  H  M  V  G  K  W  H  L  G  F  Y  R  K       155
CAG GAG GCA GGT TAT TCC ACC CAT ATG GTG GGC AAG TGG CAC CTG GGC TTC TAC CGG AAG   465

 E  C  L  P  T  R  R  G  F  D  T  F  L  G  S  L  T  G  N  V       175
GAG TGT CTG CCC ACC CGT CGG GGC TTC GAC ACC TTC CTG GGC TCG CTC ACG GGC AAT GTG   525

 D  Y  Y  T  Y  D  N  C  D  G  P  G  V  C  G  F  D  L  H  E       195
GAC TAT TAC ACC TAT GAC AAC TGT GAT GGC CCA GGC GTG TGC GGC TTC GAC CTG CAC GAG   585

 G  E  N  V  A  W  G  L  S  G  Q  Y  S  T  M  L  Y  A  Q  R       215
GGT GAG AAT GTG GCC TGG GGG CTC AGC GGC CAG TAC TCC ACT ATG CTT TAC GCC CAG CGC   645

 A  S  H  I  L  A  S  H  S  P  Q  R  P  L  F  L  Y  V  A  F       235
GCC AGC CAT ATC CTG GCC AGC CAC AGC CCT CAG CGT CCC CTC TTC CTC TAT GTG GCC TTC   705

 Q  A  V  H  T  P  L  Q  S  P  R  E  Y  L  Y  R  Y  R  T  M       255
CAG GCA GTA CAC ACA CCC CTG CAG TCC CCT CGT GAG TAC CTG TAC CGC TAC CGC ACC ATG   765

 G  N  V  A  R  R  K  Y  A  A  M  V  T  C  M  D  E  A  V  R       275
GGC AAT GTG GCC CGG CGG AAG TAC GCG GCC ATG GTG ACC TGC ATG GAT GAG GCT GTG CGC   825

 N  I  T  W  A  L  K  R  Y  G  F  Y  N  N  S  V  I  I  F  S       295
AAC ATC ACC TGG GCC CTC AAG CGC TAC GGT TTC TAC AAC AAC AGT GTC ATC ATC TTC TCC   885

 S  D  N  G  G  Q  T  F  S  G  G  S  N  W  P  L  R  G  R  K       315
AGT GAC AAT GGT GGC CAG ACT TTC TCG GGG GGC AGC AAC TGG CCG CTC CGA GGA CGC AAG   945

 G  T  Y  W  E  G  G  V  R  G  L  G  F  V  H  S  P  L  L  K       335
GGC ACT TAT TGG GAA GGT GGC GTG CGG GGC CTA GGC TTT GTC CAC AGT CCC CTG CTC AAG  1005
```

FIGURE 10A

```
 R  K  Q  R  T  S  R  A  L  M  H  I  T  D  W  Y  P  T  L  V      355
CGA AAG CAA CGG ACA AGC CGG GCA CTG ATG CAC ATC ACT GAC TGG TAC CCG ACC CTG GTG  1065

 G  L  A  G  G  T  T  S  A  A  D  G  L  D  G  Y  D  V  W  P      375
GGT CTG GCA GGT GGT ACC ACC TCA GCA GCC GAT GGG CTA GAT GGC TAC GAC GTG TGG CCG  1125

 A  I  S  E  G  R  A  S  P  R  T  E  I  L  H  N  I  D  P  L      395
GCC ATC AGC GAG GGC CGG GCC TCA CCA CGC ACG GAG ATC CTG CAC AAC ATT GAC CCA CTC  1185

 Y  N  H  A  Q  H  G  S  L  E  G  G  F  G  I  W  N  T  A  V      415
TAC AAC CAT GCC CAG CAT GGC TCC CTG GAG GGC GGC TTT GGC ATC TGG AAC ACC GCC GTG  1245

 Q  A  A  I  R  V  G  E  W  K  L  L  T  G  D  P  G  Y  G  D      435
CAG GCT GCC ATC CGC GTG GGT GAG TGG AAG CTG CTG ACA GGA GAC CCC GGC TAT GGC GAT  1305

 W  I  P  P  Q  T  L  A  T  F  P  G  S  W  W  N  L  E  R  M      455
TGG ATC CCA CCG CAG ACA CTG GCC ACC TTC CCG GGT AGC TGG TGG AAC CTG GAA CGA ATG  1365

 A  S  V  R  Q  A  V  W  L  F  N  I  S  A  D  P  Y  E  R  E      475
GCC AGT GTC CGC CAG GCC GTG TGG CTC TTC AAC ATC AGT GCT GAC CCT TAT GAA CGG GAG  1425

 D  L  A  G  Q  R  P  D  V  V  R  T  L  L  A  R  L  A  E  Y      495
GAC CTG GCT GGC CAG CGG CCT GAT GTG GTC CGC ACC CTG CTG GCT CGC CTG GCC GAA TAT  1485

 N  R  T  A  I  P  V  R  Y  P  A  E  N  P  R  A  H  P  D  F      515
AAC CGC ACA GCC ATC CCG GTA CGC TAC CCA GCT GAG AAC CCC CGG GCT CAT CCT GAC TTT  1545

 N  G  G  A  W  G  P  W  A  S  D  E  E  E  E  E  E  E  G  R      535
AAT GGG GGT GCT TGG GGG CCC TGG GCC AGT GAT GAG GAA GAG GAG GAA GAG GAA GGG AGG  1605

 A  R  S  F  S  R  G  R  R  K  K  K  C  K  I  C  K  L  R  S      555
GCT CGA AGC TTC TCC CGG GGT CGT CGC AAG AAA AAA TGC AAG ATT TGC AAG CTT CGA TCC  1665

 F  F  R  K  L  N  T  R  L  M  S  Q  R  I  *                    570
TTT TTC CGT AAA CTC AAC ACC AGG CTA ATG TCC CAA CGG ATC TGA                      1710

TGGTGGGGAGGGAGAAAACTGTCCTTTAGAGGATCTTCCCCACTCCGGCTTGGCCCTGCTGTTTCTCAGGGAGAAGCCT

GTCACATCTCCATCTACAGGGAGTTGGAGGGTGTAGAGTCCCTTGGTTGAACAGGGTAGGGAGCCTGGATAGGAGTGGG

TGGGAATAAACCAGACTGGGATGCCTGTGTCTCAGTCCTGCCTCCTCACGGACTTGCTCTGTGACCTCAGGTGACCCAC

ATGAGCTTTTAGCCTCAGTTTCCTCATCTGTAAAATGAGCTCTAATGACTTTGTGACTCTTTGGTGTGGCCCTGGAGCC

TGGGGCCACGGTGGAGTTCCTGGCCGGCCTTGCCACTTGACAACTCCTTTAAGGCTTCCCCCTTAACACGGGATCCCTG

TGGTGGTGTTTGGGAGTTGCCTGGAGGCAACTCCAAGCCTGGCCCCCAGCTGAAGCATGGCAATCTGGCTGCTCTCTAC

AGGGACCCCCAAGCGCTGTGGGTGGAGGGCAGGGGTCGGGGGGGGTTGACCTTCTTGGGTCTTCACATGGCCTAGGCCAG

TCCTCCGGTCAGACTGGTGTCAGGCACCGTGGTGCAAAATTCCTCTTCTGGCCCCTCCAGTACCCAGAGAAACTGGCTG

GGCCATTAACTGCTGCAGCACCAAGGGTGGTAGAAAGAGCTGTGAAGAGCCCCCAAACCAGTACCAGGACACCTGGGTT

CTCCTGTGACCTGGGGCACAGTTCTTGCCCTCTAGGCCTTGATTTCCCCACCTGCAAGTGGGGATGCCAGCCCTGGCTC

TGCCTCCTTCATGAGGCTCTGGAAGACTGGCCAAGGTTGTGGAGGAGCTTGTGAACTTGATTAAAGTGTCGTAACATGG

AAAAAAAAAAAAAAAAAAAAAAGGGCGG
```

FIGURE 10B

Analysis of 25278 (569 aa)

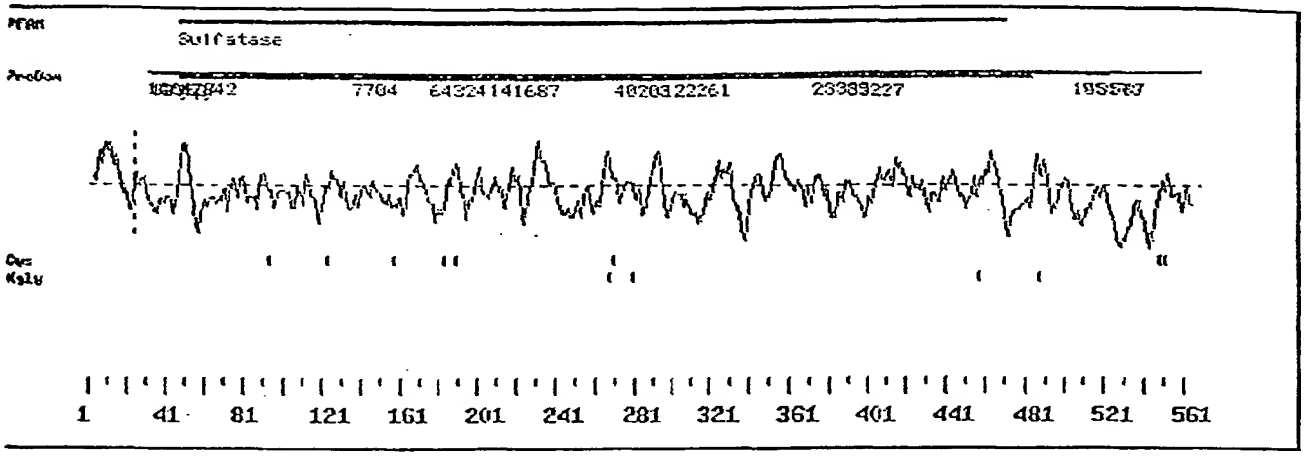

FIGURE 11

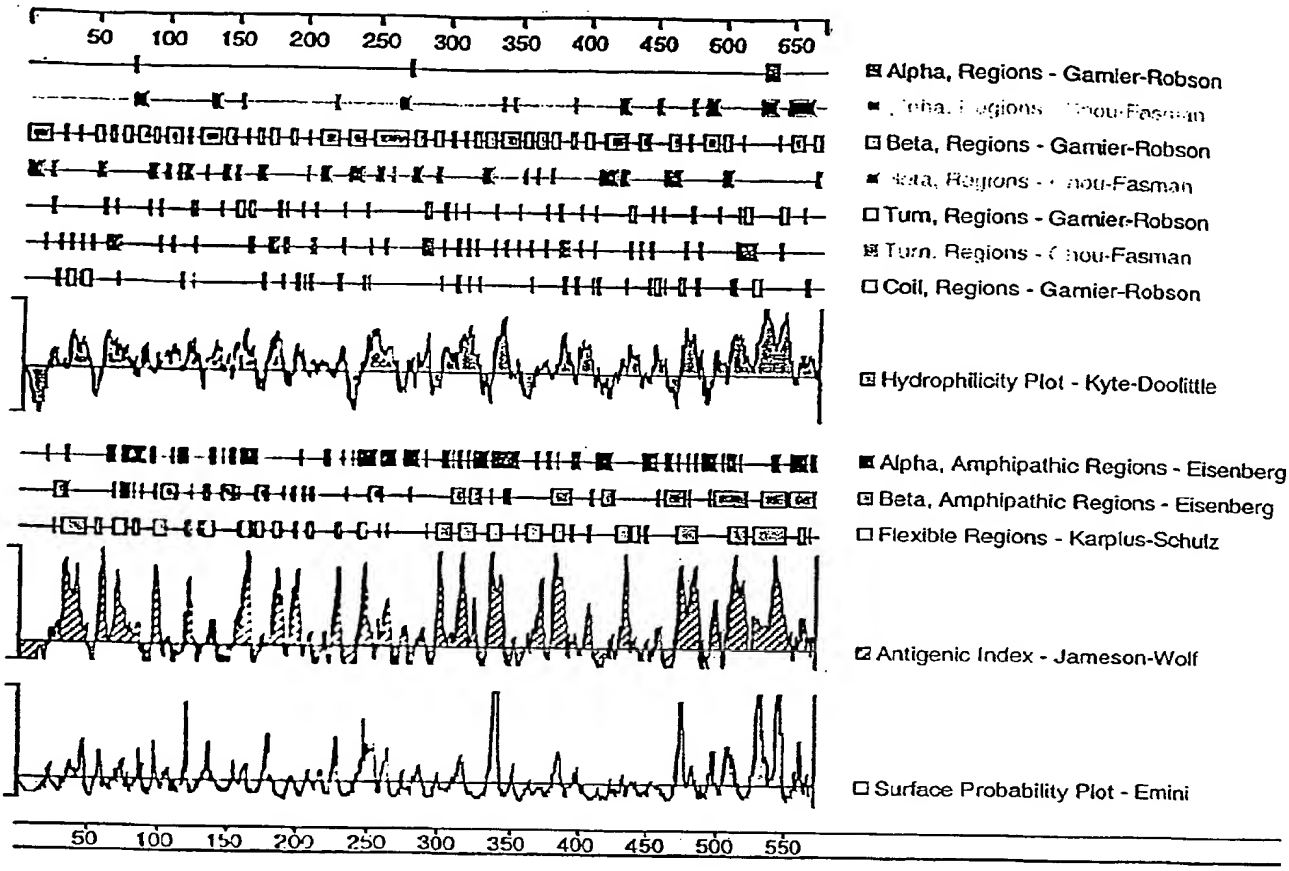

FIGURE 12

: version: Release 12.2 of February 1995

101|PDOC00001|ASN_GLYCOSYLATION N-glycosylation site.

276	NITW	279
288	NNSV	291
466	NISA	469
496	NRTA	499

104|PDOC00004|CAMP_PHOSPHO_SITE cAMP- and cGMP-dependent protein kinase phosphorylation site.

314	RKGT	317

105|PDOC00005|PKC_PHOSPHO_SITE Protein kinase C phosphorylation site.

102	TGR	104
160	TRR	162
244	SPR	246
340	TSR	342
383	SPR	385
457	SVR	459
566	SQR	568

06|PDOC00006|CK2_PHOSPHO_SITE Casein kinase II phosphorylation site.

67	SDIE	70
244	SPRE	247
268	TCMD	271
317	TYWE	320
363	SAAD	366
525	SDEE	528

07|PDOC00007|TYR_PHOSPHO_SITE Tyrosine kinase phosphorylation site.

134	KLQEAGY	140

08|PDOC00008|MYRISTYL N-myristoylation site.

110	GLQHSI	115
169	GSLTGN	174
205	GQYSTK	210
300	GQTFSG	305
321	GGVRGL	326
356	GLAGGT	361
402	GSLEGG	407
409	GIWNTA	414
447	GSWWNL	452

19|PDOC00009|AMIDATION Amidation site.

312	RGRK	315
541	RGRR	544

12|PDOC00117|SULFATASE_2 Sulfatases signature 2.

139	GYSTHMVGKW	148

23|PDOC00117|SULFATASE_1 Sulfatases signature 1.

91	PICTPSRSQLLTG	103

FIGURE 13

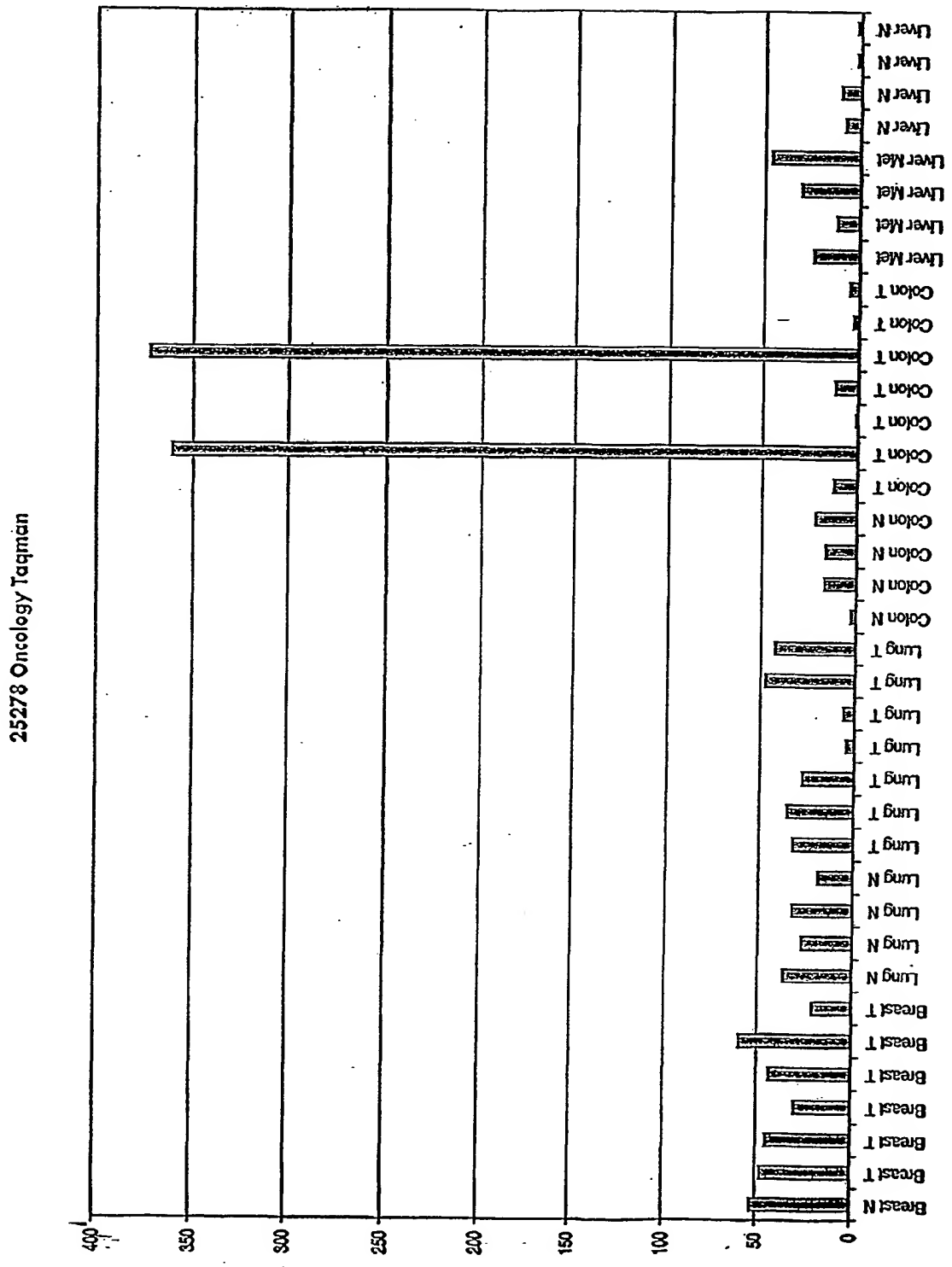

FIGURE 14.

26212 seqs

DNA sequence (nt 706-2118 coding)

CACGCGTCCGCCCACGCGTCCGTGGAGATATTAACTTTTTTCTTTTTTTTTTTCCTTGGTGGAAGCTGCTCTAGGGAGGGGGGAGGAGGA
GGAGAAAGTGAAATGTGCTGGAGAAGAGCGAGCCCTCCTTGTTCTTCCGGAGTCCCATCCATTAAGCCATCACTTCTGGAAGATTAAAGT
TGTCGGACATGGTGACAGCTGAGAGGAGAGGAGGATTTCTTGCCAGGTGGAGAGTCTTCACCGTCTGTTGGGTGCATGTGTGCGCCCGCA
GCGGCGCGGGGCGCGTGGTTCTCCGCGTGGAGTCTCACCTGGGACCTGAGTGAATGGCTCCCAGGGGCTGTGCGGGGCATCCGCCTCCGC
CTTCTCCACAGGCCTGTGTCTGTCCTGGAAAGATGCTAGCAATGGGGGCGCTGGCAGGATTCTGGATCCTCTGCCTCCTCACTTATGGTT
ACCTGTCCTGGGGCCAGGCCTTAGAAGAGGAGGAAGAAGGGGCCTTACTAGCTCAAGCTGGAGAGAAACTAGAGCCCAGCACAACTTCCA
CCTCCCAGCCCCATCTCATTTTCATCCTAGCGGATGATCAGGGATTTAGAGATGTGGGTTACCACGGATCTGAGATTAAAACACCTACTC
TTGACAAGCTCGCTGCCGAAGGAGTTAAACTGGAGAACTACTATGTCCAGCCTATTTGCACACCATCCAGGAGTCAGTTTATTACTGGAA
AGTATCAGATACACACCGGACTTCAACATTCTATCATAAGACCTACCCAACCCAACTGTTTACCTCTGGACAATGCCACCCTACCTCAGA
AACTGAAGGAGGTTGGATATTCAACGCATATGGTCGGAAAATGGCACTTGGGTTTTTACAGAAAAGAATGCATGCCCACCAGAAGAGGAT
TTGATACCTTTTTTGGTTCCCTTTTGGGAAGTGGGGATTACTATACACACTACAAATGTGACAGTCCTGGGATGTGTGGCTATGACTTGT
ATGAAAACGACAATGCTGCCTGGGACTATGACAATGGCATATACTCCACACAGATGTACACTCAGAGAGTACAGCAAATCTTAGCTTCCC
ATAACCCCACAAAGCCTATATTTTTATATATTGCCTATCAAGCTGTTCATTCACCACTGCAAGCTCCTGGCAGGTATTTCGAACACTACC
GATCCATTATCAACATAAACAGGAGGAGATATGCTGCCATGCTTTCCTGCTTAGATGAAGCAATCAACAACGTGACATTGGCTCTAAAGA
CTTATGGTTTCTATAACAACAGCATTATCATTTACTCTTCAGATAATGGTGGCCAGCCTACGGCAGGAGGGAGTAACTGGCCTCTCAGAG
GTAGCAAAGGAACATATTGGGAAGGAGGGATCCGGGCTGTAGGCTTTGTGCATAGCCCACTTCTGAAAAACAAGGGAACAGTGTGTAAGG
AACTTGTGCACATCACTGACTGGTACCCCACTCTCATTTCACTGGCTGAAGGACAGATTGATGAGGACATTCAACTAGATGGCTATGATA
TCTGGGAGACCATAAGTGAGGGTCTTCGCTCACCCCGAGTAGATATTTTGCATAACATTGACCCCATATACACCAAGGCAAAAAATGGCT
CCTGGGCAGCAGGCTATGGGATCTGGAACACTGCAATCCAGTCAGCCATCAGAGTGCAGCACTGGAAATTGCTTACAGGAAATCCTGGCT
ACAGCGACTGGGTCCCCCCTCAGTCTTTCAGCAACCTGGGACCGAACCGGTGGCACAATGAACGGATCACCTTGTCAACTGGCAAAAGTG
TATGGCTTTTCAACATCACAGCCGACCCATATGAGAGGGTGGACCTATCTAACAGGTATCCAGGAATCGTGAAGAAGCTCCTACGGAGGC
TCTCACAGTTCAACAAAACTGCAGTGCCGGTCAGGTATCCCCCCAAAGACCCCAGAAGTAACCCTAGGCTCAATGGAGGGGTCTGGGGAC
CATGGTATAAAGAGGAAACCAAGAAAAAGAAGCCAAGCAAAAATCAGGCTGAGAAAAAGCAAAAGAAAAGCAAAAAAAAGAAGAAGAAAC
AGCAGAAAGCAGTCTCAGGTTCAACTTGCCATTCAGGTGTTACTTGTGGATAAGCACAAATATTTCCTGTTTGGTTAAACTTTAATCAGT
TCTTATCTTTCATCTGTTTCCTAGGTAAACCAGCAAATTTGGCTCGATAATATCGCTGGCCTAAGCGTCAGGCTTGTTTTCATGCTGTGC
CAC

Protein sequence

MAPRGCAGHPPPPSPQACVCPGKMLAMGALAGFWILCLLTYGYLSWGQALEEEEEGALLAQAGEKLEPSTTSTSQPHLIFILADDQGFRD
VGYHGSEIKTPTLDKLAAEGVKLENYYVQPICTPSRSQFITGKYQIHTGLQHSIIRPTQPNCLPLDNATLPQKLKEVGYSTHMVGKWHLG
FYRKECMPTRRGFDTFFGSLLGSGDYYTHYKCDSPGMCGYDLYENDNAAWDYDNGIYSTQMYTQRVQQILASHNPTKPIFLYIAYQAVHS
PLQAPGRYFEHYRSIININRRRYAAMLSCLDEAINNVTLALKTYGFYNNSIIIYSSDNGGQPTAGGSNWPLRGSKGTYWEGGIRAVGFVH
SPLLKNKGTVCKELVHITDWYPTLISLAEGQIDEDIQLDGYDIWETISEGLRSPRVDILHNIDPIYTKAKNGSWAAGYGIWNTAIQSAIR
VQHWKLLTGNPGYSDWVPPQSFSNLGPNRWHNERITLSTGKSVWLFNITADPYERVDLSNRYPGIVKKLLRRLSQFNKTAVPVRYPPKDP
RSNPRLNGGVWGPWYKEETKKKKPSKNQAEKKQKKSKKKKKQQKAVSGSTCHSGVTCG

FIGURE 15

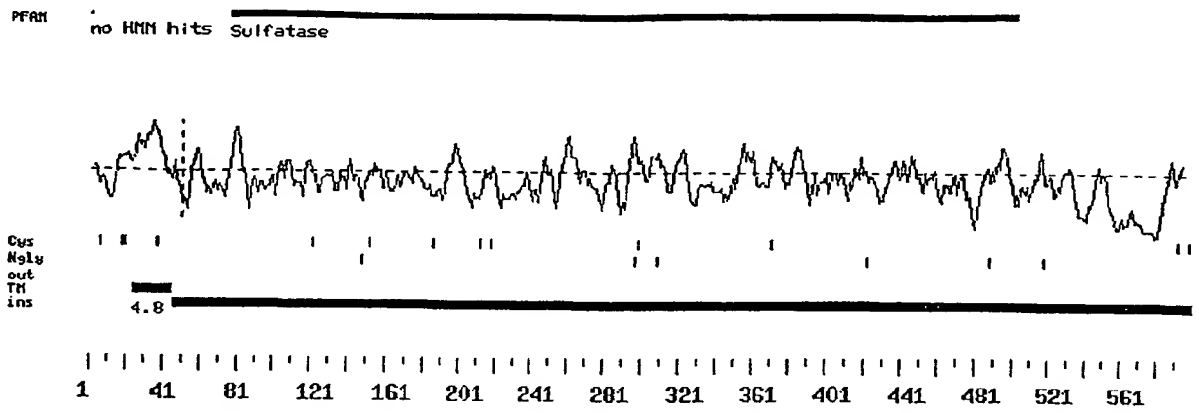

FIGURE 16

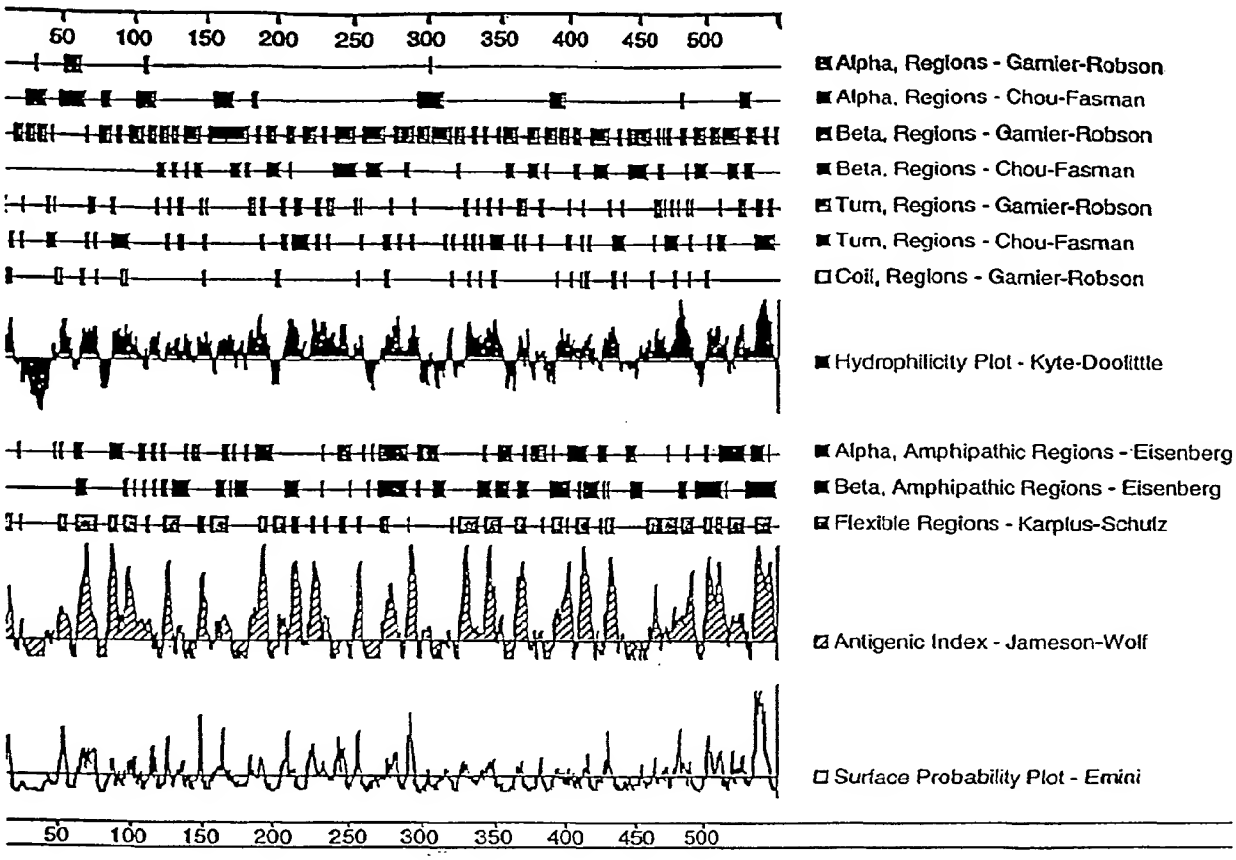

FIGURE 17

Prosite Pattern Matches for 26212

Prosite version: Release 12.2 of February 1995

>PS00001|PDOC00001|ASN_GLYCOSYLATION N-glycosylation site.

Query:	157	NATL	160
Query:	306	NVTL	309
Query:	318	NNSI	321
Query:	431	NGSW	434
Query:	497	NITA	500
Query:	527	NKTA	530

>PS00004|PDOC00004|CAMP_PHOSPHO_SITE cAMP- and cGMP-dependent protein kinase phosphorylation site.

Query:	521	RRLS	524
Query:	562	KKPS	565

>PS00005|PDOC00005|PKC_PHOSPHO_SITE Protein kinase C phosphorylation site.

Query:	131	TGK	133
Query:	189	TRR	191
Query:	243	TQR	245
Query:	413	SPR	415
Query:	489	TGK	491
Query:	509	SNR	511
Query:	559	TKK	561
Query:	576	SKK	578

>PS00006|PDOC00006|CK2_PHOSPHO_SITE Casein kinase II phosphorylation site.

FIGURE 18A

```
Query:  298     SCLD    301

Query:  347     TYWE    350

Query:  386     SLAE    389

Query:  406     TISE    409

>PS00007|PDOC00007|TYR_PHOSPHO_SITE Tyrosine kinase
phosphorylation site.

Query:  163     KLKEVGY 169

>PS00008|PDOC00008|MYRISTYL N-myristoylation site.

Query:  28      GALAGF  33

Query:  56      GALLAQ  61

Query:  139     GLQHSI  144

Query:  198     GSLLGS  203

Query:  235     GIYSTQ  240

Query:  329     GGQPTA  334

Query:  343     GSKGTY  348

Query:  351     GGIRAV  356

Query:  432     GSWAAG  437

Query:  439     GIWNTA  444

>PS00149|PDOC00117|SULFATASE_2 Sulfatases signature 2.

Query:  168     GYSTHMVGKW      177

>PS00523|PDOC00117|SULFATASE_1 Sulfatases signature 1.

Query:  120     PICTPSRSQFITG   132
```

FIGURE 18B

```
Alignments of top-scoring domains:
Sulfatase: domain 1 of 1, from 36 to 462: score 323.0, E = 3.5e-93
                  *->PNillilaDDlGigdlGcyGnptirtpniDrLAeeGlrFtnayvttp
                     PN+++ilaDD+G+gdlG+  + t  t n+D +A+eG+rF ++ ++++
       25277    36    PNFVIILADDMGWGDLGANWAETKDTANLDKMASEGMRFVDFHAAAS 82

                  lCtPSRAalLTGRyphrtGmytnnragvlpftgwsleGglpldettlpel
                   C+PSRA+lLTGR+  r+G++ n  +        +s +Gglpl+ettl+e+
       25277    83 TCSPSRASLLTGRLGLRNGVTRNFAV-------TS-VGGLPLNETTLAEV 124

                  LkeaGYaTgmvGKWHlgyneessasdfahlPlgrGFdyfygnlGGEdQWY
                  L++aGY+Tg++GKWHlg++ ++        +P  rGFdy++g
       25277    125 LQQAGYVTGIIGKWHLGHHGSY-------HPNPRGFDYYFG--------- 158

                  plvdallpfttndtytceggygfskdvalkplgalgvneveapdkaladyk
                        +p+++ ++ c                                  +
       25277    159 ------IPYSH-DMGCT---------------------------------D 169

                  tagalnvphhvfEWadryagavdvgrpflavlifprpaacflypnatvvs
                  t+g+ + p          + ++++ +r    + ++    + a+ ly n+ +v+
       25277    170 TPGYNHPP------CPACPQGDGPSRNLQRDCY--TDVALPLYENLNIVE 211

                  qpmphspltaPrpwqlladealpflerngqrdkpfflylsykhvhiprda
                  qp   s l+      q +a++a +f++r+ +   +pf+ly++++h+h+p
       25277    212 QPVNLSSLA-----QKYAEKATQFIQRASTSGRPFLLYVALAHMHVP--- 253

                  pmlfsskdfagssrrglYglilDsveemDdgvgrvlnaLdelNGlldnTl
                    l+  +  a    r lYg   + + emD +vg++ + +d    + +nT+
       25277    254 --LPVTQLPAAPRGRSLYG---AGLWEMDSLVGQIKDKVDHT--VKENTF 296

                  iiFTSllDhGghlgahghlgiragGsngpfrg..........gKgtnlye
                    FT   D+G+ ++ + +      Gs gpf g  +++++++++K+t+ +e
       25277    297 LWFTG--DNGPWAQKCELA-----GSVGPFTGfwqtrqggspAKQTT-WE 338

                  gGtRvPlivrwPeGiiapqqvsdelvslmDlfPTildLAGaplPgvaagv
                  gG+RvP++++wP G+ + + +s +l s++D+fPT+++LA a+lP
       25277    339 GGHRVPALAYWP-GRVPVNVTSTALLSVLDIFPTVVALAQASLP------ 381

                  kdrilDGvsLlplLlgaagssrhetlfyesycnegrgflpavrwgkkkah
                  + r  DGv++ ++L g+  +++h  lf++   n g     a++ +
       25277    382 QGRRFDGVDVSEVLFGR-SQPGHRVLFHP---NSG-----AAGEFGALQT 422

                  frtpniagwqrvdfddvwklfntvedfnrsgddacrhgdvckclgkprrs
                  +r +           + k+f++++       ++++  g+  +  +
       25277    423 VRLE------------RYKAFYITGGAR---ACDGSTGFELQHKF----- 452

                  vthhdppllydlsrDP<-*
                        pl ++l+ D
       25277    453 ------PLIFNLEDDT    462

//
```

FIGURE 19

```
Alignments of top-scoring domains:
Sulfatase: domain 1 of 1, from 43 to 467: score 268.9, E = 6.5e-77
                 *->PNillilaDDlGigdlGcyGnptirtpniDrLAeeGlrFtnayvttp
                    PNi+l+l+DD++ ++lG+     ++ ++   +  +G  F na+vttp
       23553    43 PNIILVLTDDQD-VELGSLQ---VMNKTRKIMEHGGATFINAFVTTP 85

                    lCtPSRAalLTGRyphrtGnytnnragvlpftgwsleGglpldettlpel
                    +C+PSR++ LTG+y h++++ytnn++  ++++ w+       ++ +t++++
       23553    86 MCCPSRSSMLTGKYVHNHNVYTNNEN--CSSPSWQ----AMHEPRTFAVY 129

                    LkeaGYaTgmvGKWHlgyneessasdfahlPlgrG.FdyfygnlGGEdQW
                    L + GY+T+++GK++++yn ++        +P+g+ ++     +n
       23553   130 LNNTGYRTAFFGKYLNEYNGSY-------IPPGWReWLGLIKN------- 165

                    Yplvdallpftndtytceggygfskdvalkplgalgvneveapdkalady
                          ++f+n +  c++g           + ++++ ++++       dy
       23553   166 --------SRFYN-YTVCRNG-----------IKEKHGFDYAK------DY 190

                    ktagalnvphhvfEWadryagavdvgrpflavlifprpaacflypnatvv
                    +t++++n +          y++++                 p++++++     +
       23553   191 FTDLITNES-------INYFKMSK----------RMYPHRPVMMV----I 219

                    sqpmphspltaPrpwqlladealpflerngqrdkpfflylcykhvhiprd
                    s+ +ph p + +    ++++ +  p+ + +  +    + +++ +kh+ ++++
       23553   220 SHAAPHGPED-S---APQFSKLYPNASQH-ITPSYNYAPNMDKHWIMQYT 264

                    apmlfsskdfagssrrglYqlilDsveemDdgvgrvlnaLdelNGlldnT
                    +pml+ + +f+  ++r++ +     +++++Dd+v+r++n L e  G+l+nT
       23553   265 GPMLPIHMEFTNILQRKRLQ----TLMSVDDSVERLYNMLVET-GELENT 309

                    liiFTSllDhGghlgahghlgiragGsngpfrggKgtnlyegGtRvPliv
                    +ii+T+  DhG+h+g++g+              + gK+++ y++++RvP+++
       23553   310 YIIYTA--DHGYHIGQFGLV-----------K-GKSHP-YDFDIRVPFFI 344

                    rwPeGiiapgqvsdelvslmDlfPTildLAGaplPgvaagvkdrilDGvs
                    r+P    +pg+++ ++v ++Dl+PTild+AG++ P           +DG+s
       23553   345 RGP--SVEPGSIVPQIVLNIDLAPTILDIAGLDTP--------PDVDGKS 384

                    LlplLlgaagssrhetlfyesycnegrgflpavrwgkkkahfrtpniagw
                    +l+lL+ +     ++ ++f + + +++    +++ +    +  +f
       23553   385 VLKLLDPE----KPGNRFRT-NKKAK---IWRDTFLVERGKP-------- 418

                    qrvdfddvwklfntvedfnrsgddacrhgdvckclgkprrsvthhdppll
                          + k + +  + ++S++ + + +    +c ++++ ++ +++p +
       23553   419 -------LRKKEPSSKNICQSNHLPKYERVKELCQQARYQTA-CEQPGQK 460

                    ydlsrDP<-*
                       +D
       23553   461 WQCIEDT    467
```

FIGURE 20

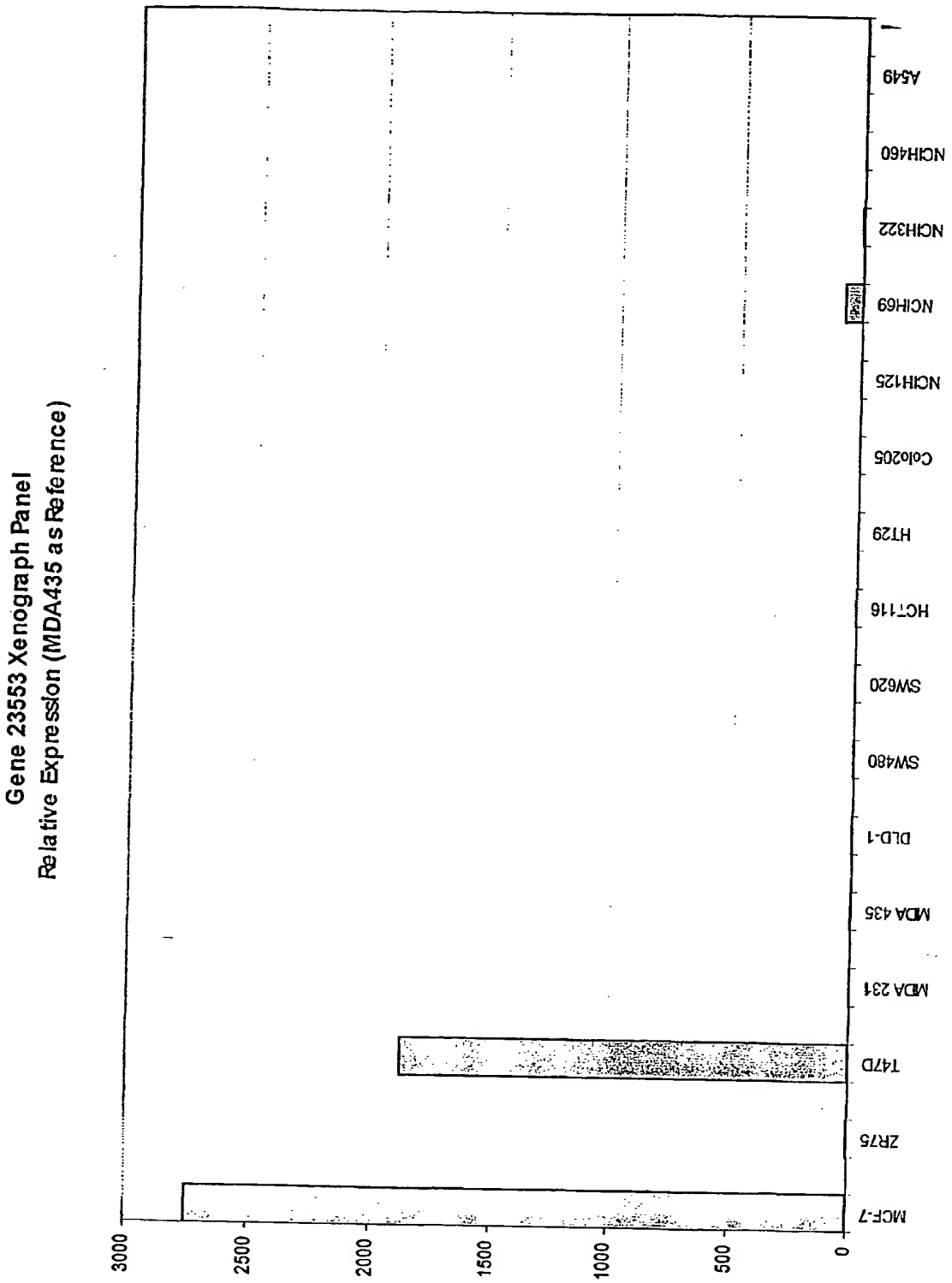

FIGURE 21

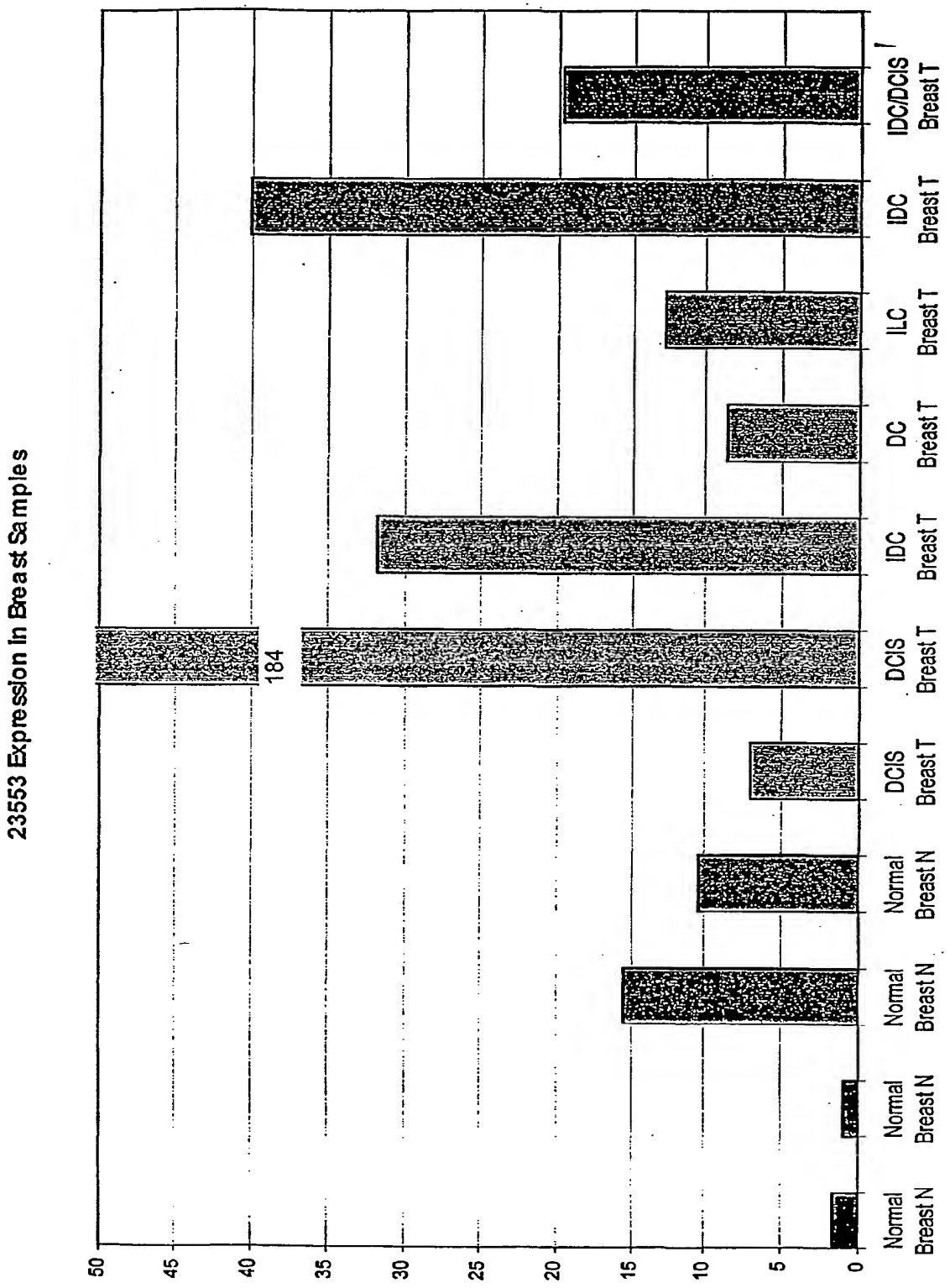

FIGURE 22

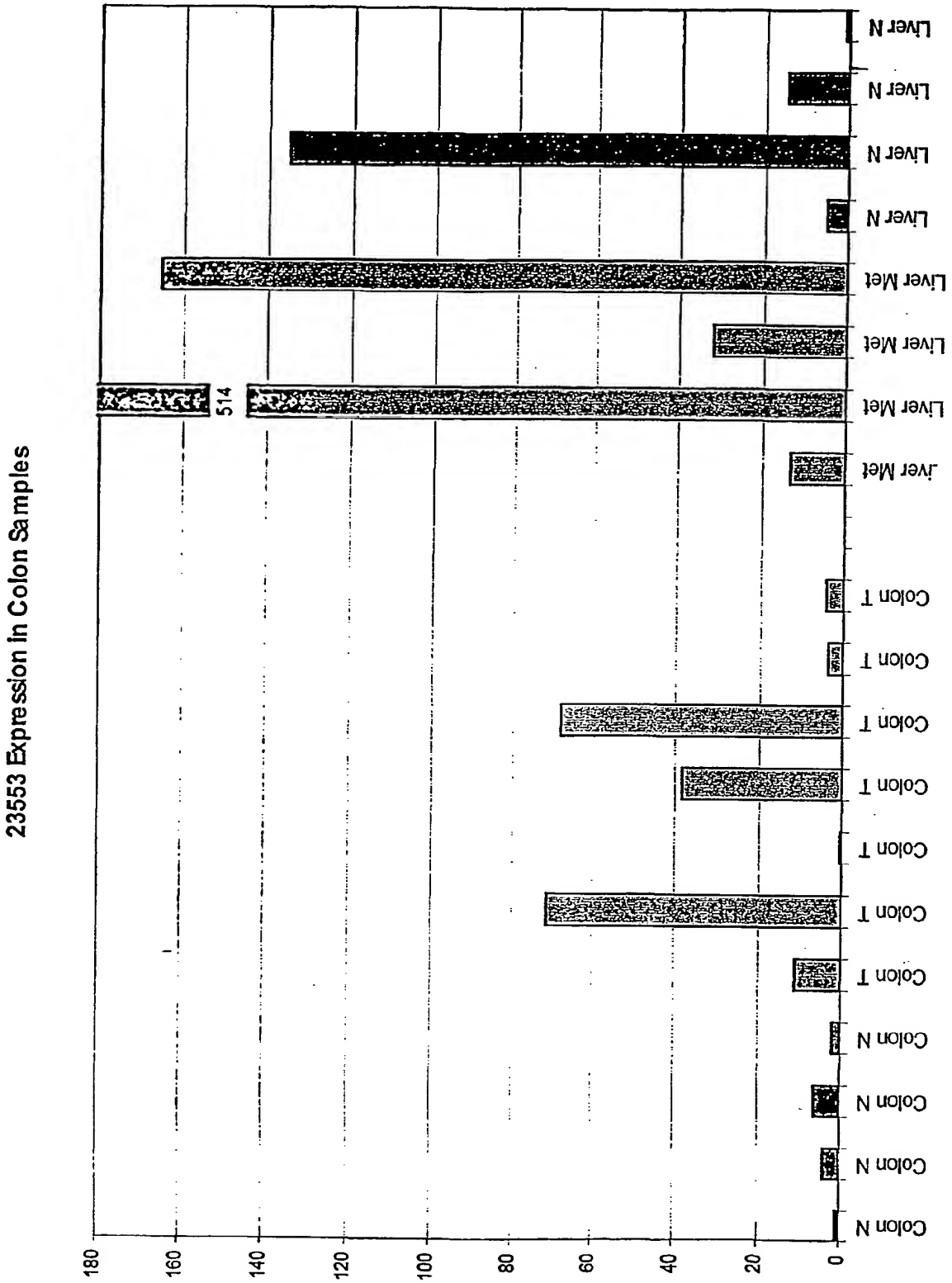

FIGURE 23

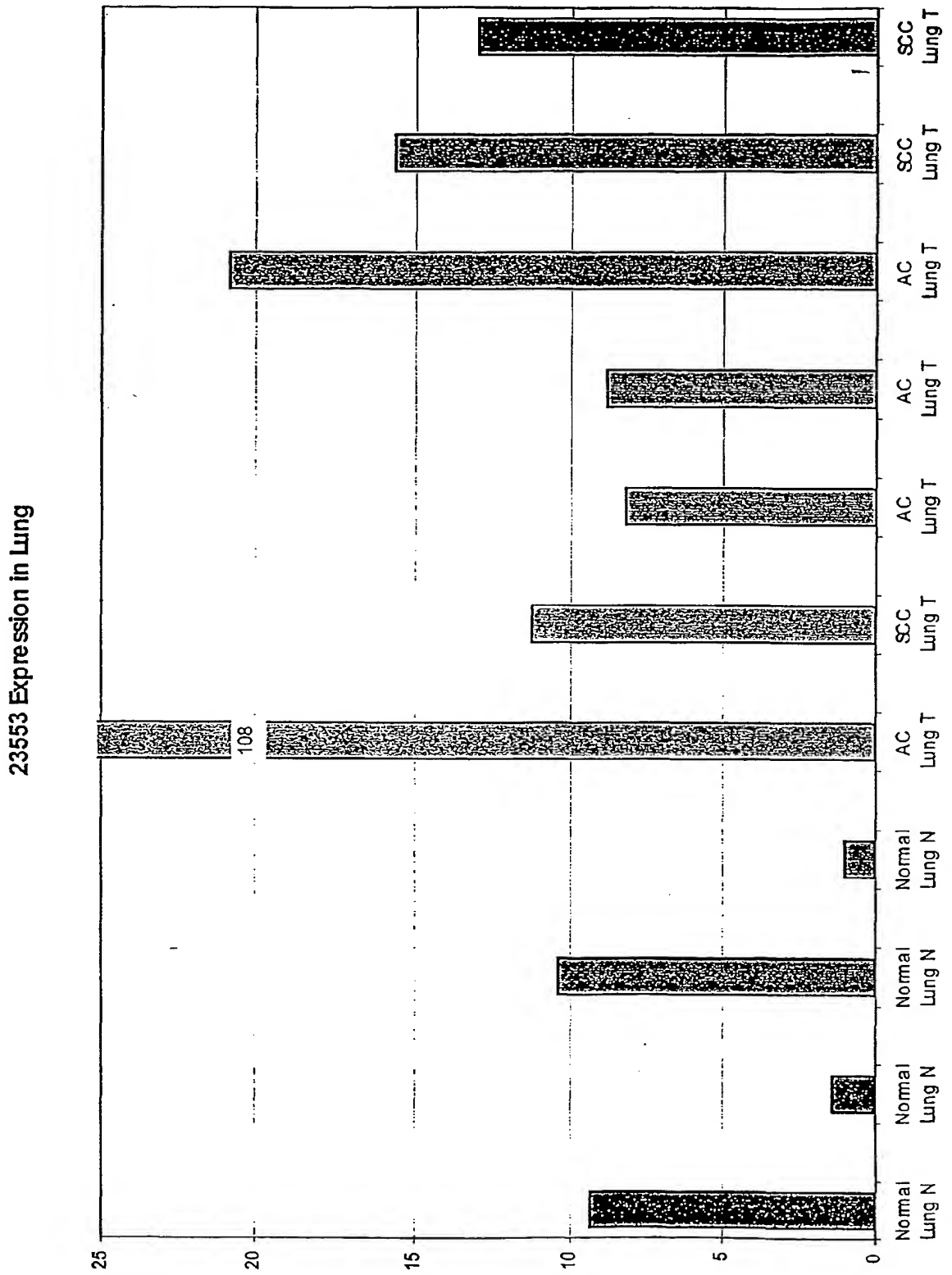

FIGURE 24

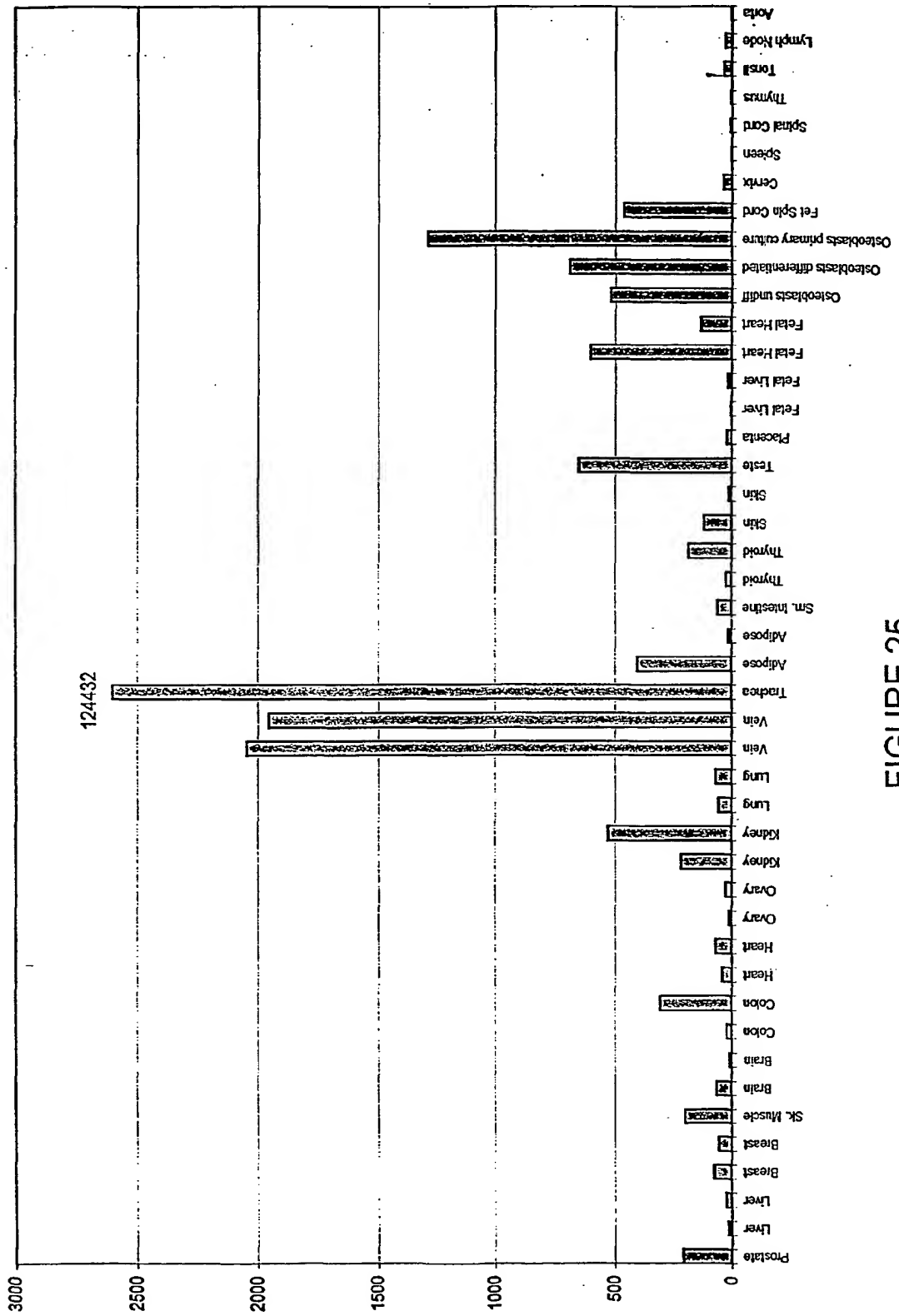

FIGURE 25

TAQMAN RESULTS

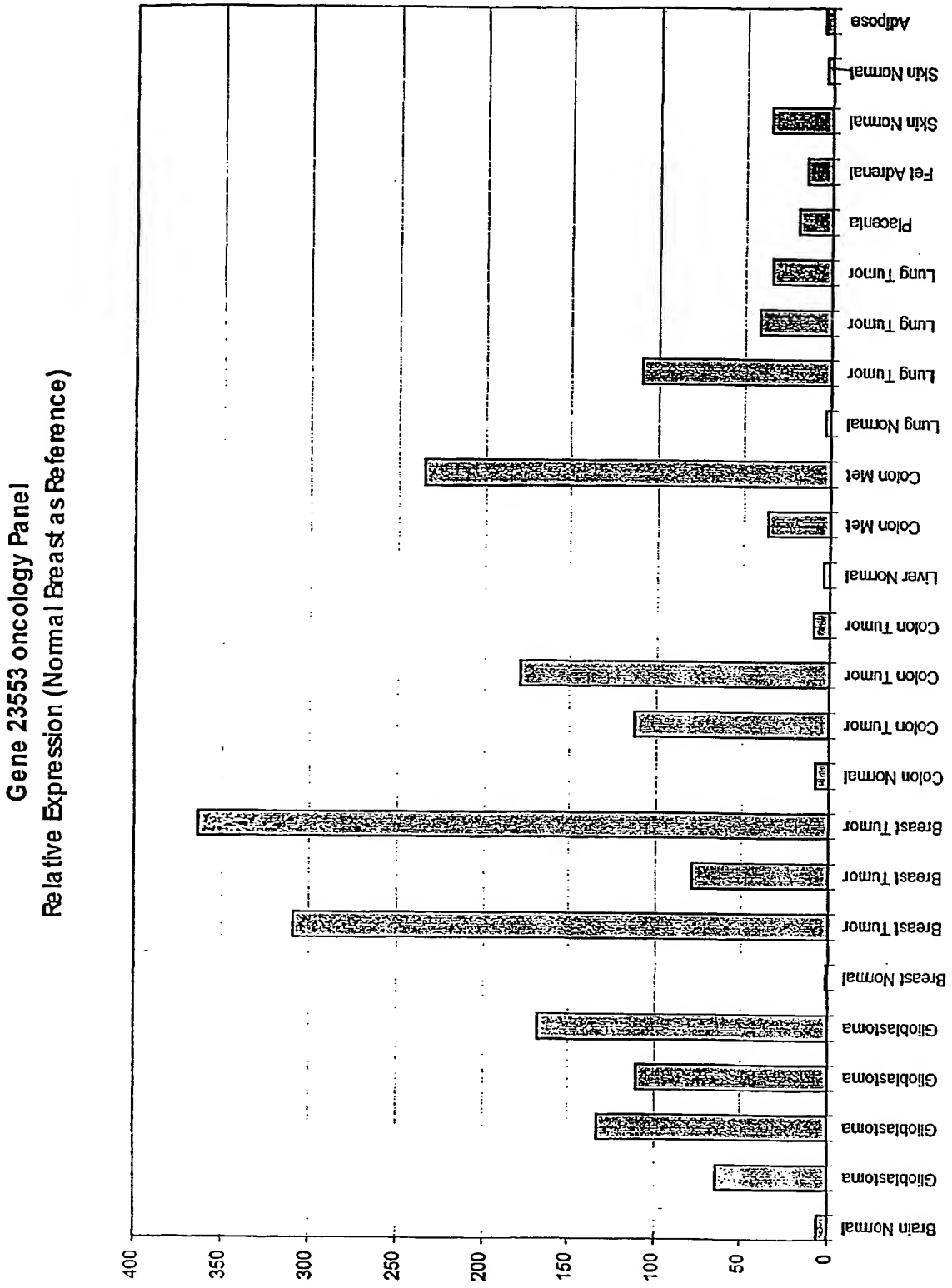

FIGURE 26

```
Alignments of top-scoring domains:
Sulfatase: domain 1 of 1, from 47 to 471: score 289.7, E = 3.6e-83
                  *->PNillilaDDlGigdlGcyGnptirtpniDrLAeeGlrFtnayvttp
                     P+i++il+DD+G+ d+G +G + i+tp++DrLA+ G+++ n y+ +p
       25278    47    PHIIFILTDDQGYHDVGYHG-SDIETPTLDRLAAKGVKLEN-YYIQP 91

                  lCtPSRAalLTGRyphrtGmytnnragvlpftgwsleGglpldettlpel
                  +CtPSR++lLTGRy+++tG+++ +      p+++      +lpld +tlp+
       25278    92 ICTPSRSQLLTGRYQIHTGLQHSIIR---PQQPN----CLPLDQVTLPQK 134

                  LkeaGYaTgmvGKWHlgyneessasdfahlPlgrGFdyfygnlGGEdQWY
                  L+eaGY T+mvGKWHlg +++++       lP++rGFd+f+g+
       25278   135 LQEAGYSTHMVGKWHLGFYRKEC------LPTRRGFDTFLGS-------- 170

                  plvdallpftndtytceggygfskdvalkplgalgvneveapdkaladyk
                        l +  d+yt+++                              ++
       25278   171 ------LTGNVDYYTYDN------------------------------CD 184

                  tagalnvphhvfEWadryagavdvgrpflavlifprpaacflypnatvvs
                   +g+ +                +d                  + + +++
       25278   185 GPGVCG---------------FD-----------------LHEGENVAWG 202

                  qpmphspltaPrpwqlladealpflcrngqrdkpfflylsykhvhiprd.
                     ++s++        +a++a    l ++     +p fly+++++vh+p ++
       25278   203 LSGQYSTML-------YAQRASHILASH-SPQRPLFLYVAFQAVHTPLQs 244

                  apmlfsskdfagssrrglYglilDsveemDdgvgrvlnaLdelNGlldnT
                  + +++  ++  g+  r+ Y+   ++v  mD++v ++  aL++  G ++n
       25278   245 PREYLYRYRTHGNVARRKYA---AMVTCMDEAVRNITWALKRY-GFYNNS 290

                  liiFTSllDhGghlgahghlgiragGsngpfrggKgtnlyegGtRvPliv
                  +iiP+S  D+Gg++           gGsn+p+rg+Kgt  +egG+R  ++v
       25278   291 VIIFSS--DNGGQTF----S----GGSNWPLRGRKGTY-WEGGVRGLGFV 329

                  rwPeGiiapgqvsdelvslmDlfPTildLAGaplPgvaagvkdrilDGvs
                  ++P   +++ ++s++l ++ D++PT++ LAG++          + lDG++
       25278   330 HSP-LLKRKQRTSRALMHITDWYPTLVGLAGGTTS------AADGLDGYD 372

                  LlplLlgaagssrhetlfye......sycnegrgflpavrwgkkkahfrt
                  ++p++ ++ +s+r e+l+++++  ++ +++   g     + g ++  + +
       25278   373 VWPAISEGRASPRTEILHNIdplynhAQHGSLEG-----GPGIWNTAVQA 417

                  pni.agwqrvdfddvwklfntvedfnrsgddacrhgdvckclgkprrsvt
                  +   + w       + +    ++d+ +++  a + g +  +       ++
       25278   418 AIRvGEWK------LLTGDPGYGDWIPPQTLATFPGSWWNLER----MAS 457

                  hhdppllydlsrDP<-*
                    +   l+++s+DP
       25278   458 VRQAVWLFNISADP    471
```

FIGURE 27

			Relative Expression
NDR 19	Breast T	DCIS	47.84
MDA 138	Breast N	Normal	52.89
NDR 01	Breast T	IDC	44.79
NDR 15	Breast T	DC	29.55
NDR 133	Breast T	ILC	43.26
MDA 161	Breast T	IDC	60.13
MDA 155	Breast T	IDC/DCIS	20.11
PIT 270	Lung N	Normal	36.00
CHT 427	Lung N	Normal	26.54
PIT 241	Lung N	Normal	31.45
PIT 298	Lung N	Normal	17.57
CHT 800	Lung T	AC	31.45
CHT 335	Lung T	SCC	35.02
CHT 447	Lung T	AC	27.19
CHT 752	Lung T	AC	3.89
CHT 799	Lung T	AC	5.74
CHT 813	Lung T	SCC	47.18
CHT 369	Lung T	SCC	42.37
CHT 371	Colon N	Normal	2.37
CHT 396	Colon N	Normal	16.34
CHT 398	Colon N	Normal	15.24
NDR 104	Colon N	Normal	20.89
CHT 520	Colon T	Adeno	11.71
CHT 122	Colon T	Adeno	360.79
CHT 536	Colon T	Adeno	1.00

FIGURE 28A

			Relative Expression
CHT 528	Colon T	Adeno	11.63
CHT 386	Colon T	Adeno	372.22
CHT 372	Colon T	Adeno	2.39
CHT 532	Colon T	Adeno	4.45
CHT 77	Liver Met	Met	23.43
CHT 321	Liver Met	Met	11.35
CHT 84	Liver Met	Met	30.38
NDR 100	Liver Met	Met	46.21
NDR 154	Liver N	Normal	7.31
CHT 322	Liver N	Normal	9.38
PIT 51	Liver N	Normal	1.77
CHT 339	Liver	Normal	1.58
PIT 265	Breast N	Normal	37.40
MDA 335	Breast N	Normal	45.57
NDR 132	Breast T	DCIS	19.56
NDR 13	Breast N	Normal	6.73
NDR 56	Breast N	Normal	20.61

FIGURE 28B

```
uments of top-scoring domains:
tase: domain 1 of 1, from 76 to 502: score 324.5, E = 1.3e-93
             *->PNvllilaDDlGigdlgcyghptirTPnldrLAeeGlrFtnhytatp
                P+ ++ilaDD+G+ d+g++g ++i TP+ld+LA+eG+++ n+y+ +p
  26212     76 PHLIFILADDQGFRDVGYHG-SEIKTPTLDKLAAEGVKLENYYV-QP 120

             lCsPSRAaLlTGryphrhGmvsngrlgvlgftaksgglpldettLpelLk
             +C+PSR+++ TG+y++++G  +      +  + ++ +lpld +tLp+ Lk
  26212    121 ICTPSRSQFITGKYQIHTGLQH-----SIIRPTQPNCLPLDNATLPQKLK 165

             eaGYaTglvGKWHlglnensdaagdgehlPlgwrGfdyfdgflygspfty
             e GY T++vGKWHlg+++        +e+ P++ rGfd f+g l+gs ++y
  26212    166 EVGYSTHMVGKWHLGFYR-------KECMPTR-RGFDTFFGSLLGSGDYY 207

             deencdngegteppeaypeqgwlpqilgyyltdlladkalglldvasaag
              ++ cd      +p+                          ++++l+     aa
  26212    208 THYKCD-----SPGM----------------------CGYDLYENDNAA- 229

             rllakalaasrPFflyisppaphfsilfrnfkevaqpyrapqltqlfvde
                                 ++++                + ++tq+++++
  26212    230 ---------------------WDYD-------------NGIYSTQMYTQR 245

             aadfiernk.ekPfflylaflrlhvhtplfspaedleskdflgrsqrgrY
                 ++++++ kP fly a++  +vh pl++p  + e+++       r+rY
  26212    246 VQQILASHNpTKPIFLYIAYQ--AVHSPLQAPGRYFEHYRSIININRRRY 293

             gdlveemDdlvGrvldaLedlGlldNTlvifTSDnGahlegtpewygggn
             +++++  D+++++v  aL+  G ++N ++i++SDnG    g+p+ +gg+n
  26212    294 AAMLSCLDEAINNVTLALKTYGFYNNSIIIYSSDNG----GQPT-AGGSN 338

             gplkggKgygslyeGgiRvPllvrwPggiapagrvkekselvshvDlaPT
             +pl+g Kg+   +eGgiR  ++v++P   + +g+v +  elv++ D++PT
  26212    339 WPLRGSKGTY--WEGGIRAVGFVHSP-LLKNKGTVCK--ELVHITDWYPT 383

             ildlAGaplPkvanGakdrplDGvsllplllggaapsrrahetlfhyngk
             + +lA +  ++      d  lDG++++ +  +g + s+  + +++h+
  26212    384 LISLAEGQIDE------DIQLDGYDIWETISEGLR-SP--RVDILHN--- 421

             grklravrwprksgktpklkahfftpaf.....................
                ++ ++  +k+   + + a +  ++   ++  + ++ +  +++++ ++
  26212    422 ---IDPIYTKAKN---GSWAAGYGIWNTaiqsairvqhwkllltgnpgysd 465

             ....dddtnngwecvgtvsqaddiedcrcegvetvthhdppelyDlsrDP
               ++++  n+g             +  ++ e   t+   +  +l++ ++DP
  26212    466 wvppQSFSNLG-----------PNRWHNER-ITLSTGKSVWLFNITADP 502

             <-*

  26212      -     -
```

FIGURE 29

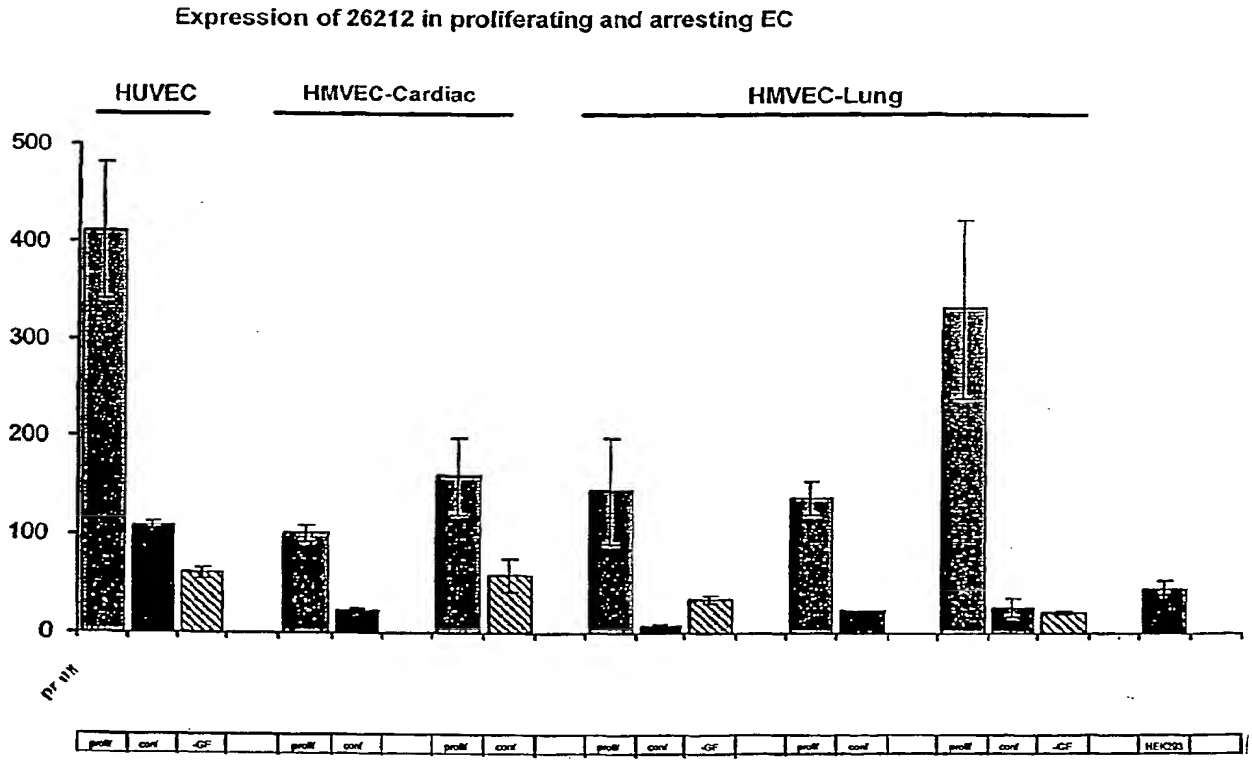

FIGURE 30

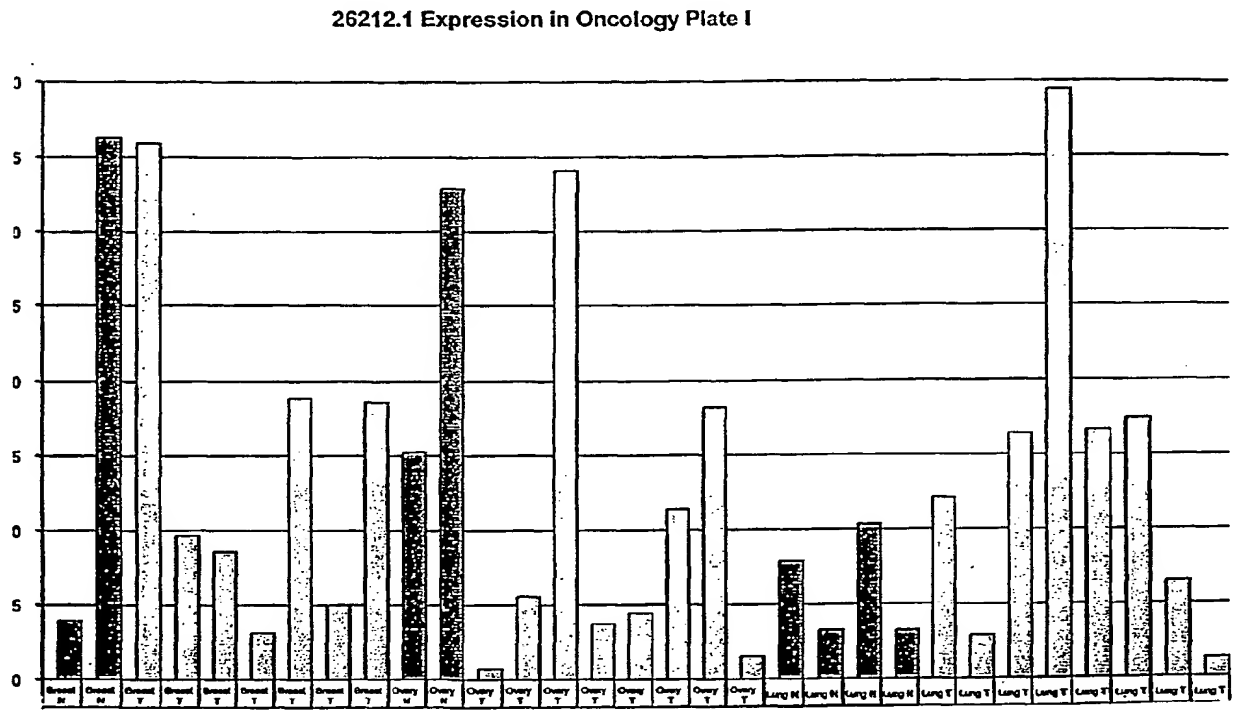

FIGURE 31A

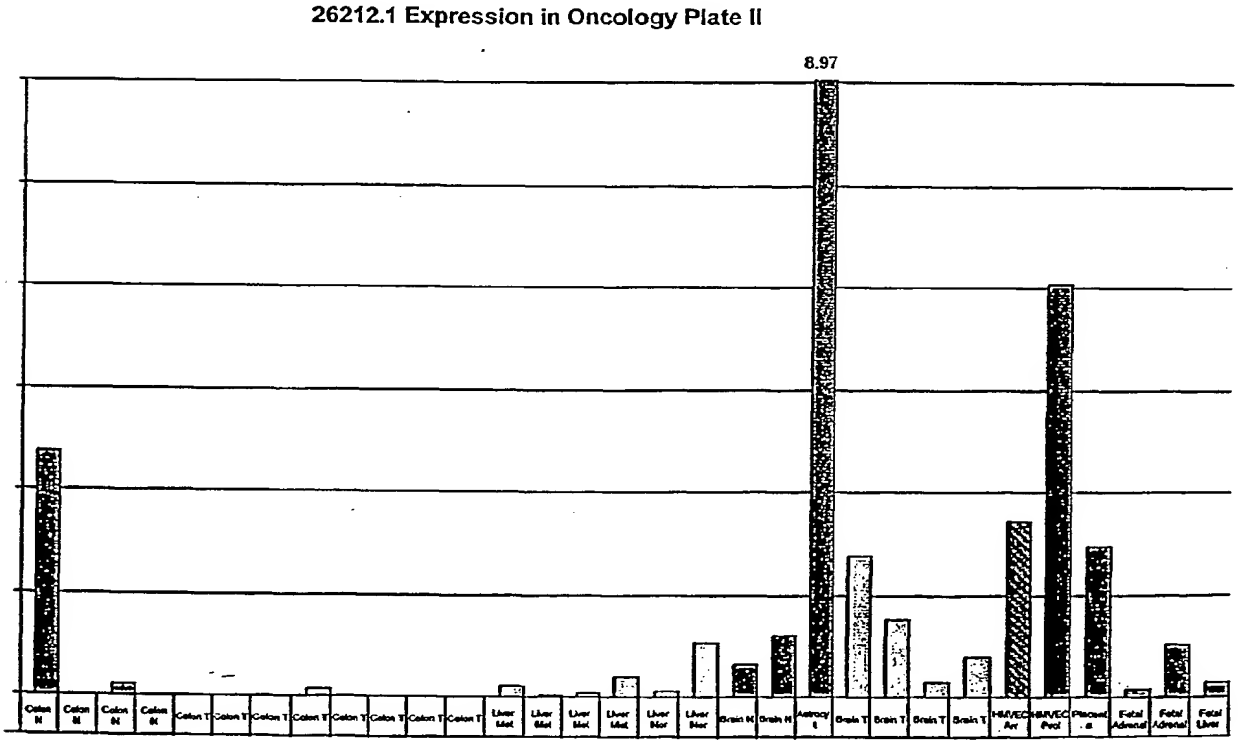

FIGURE 31B

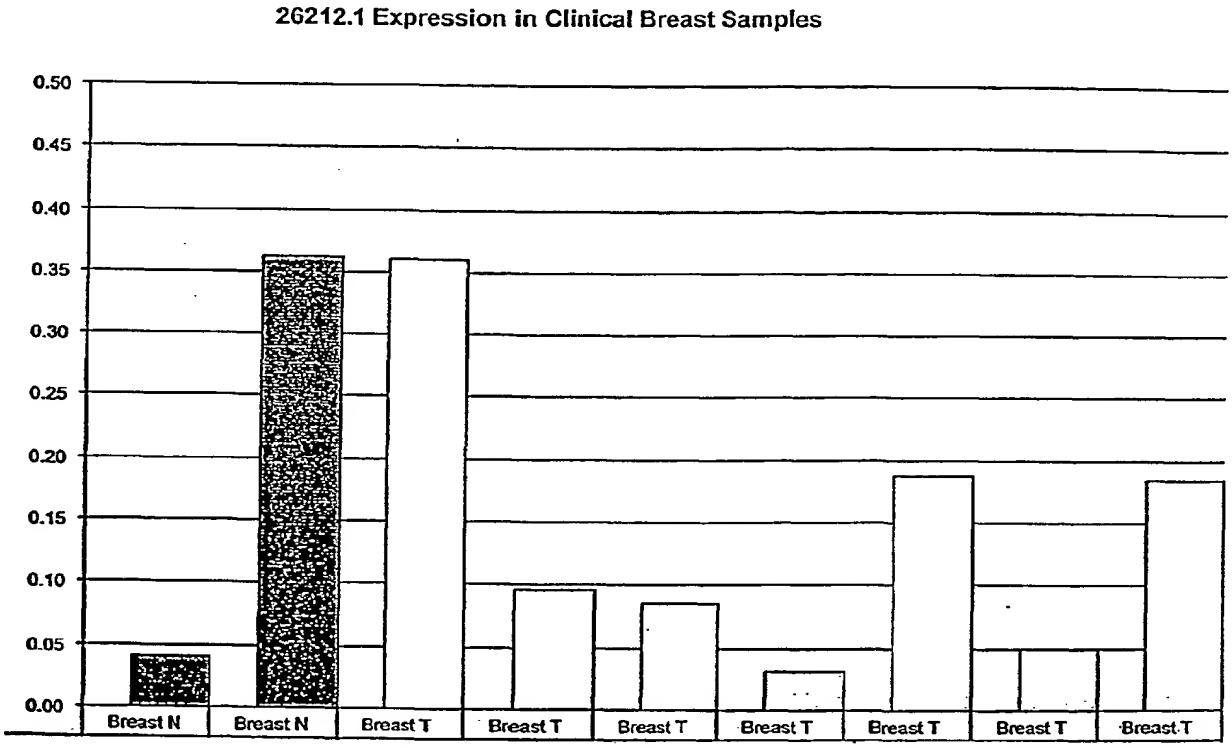
26212.1 Expression in Clinical Breast Samples
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Breast N
Breast N
Breast T
Breast T
Breast T
Breast T
Breast T
Breast T
Breast T

FIGURE 32

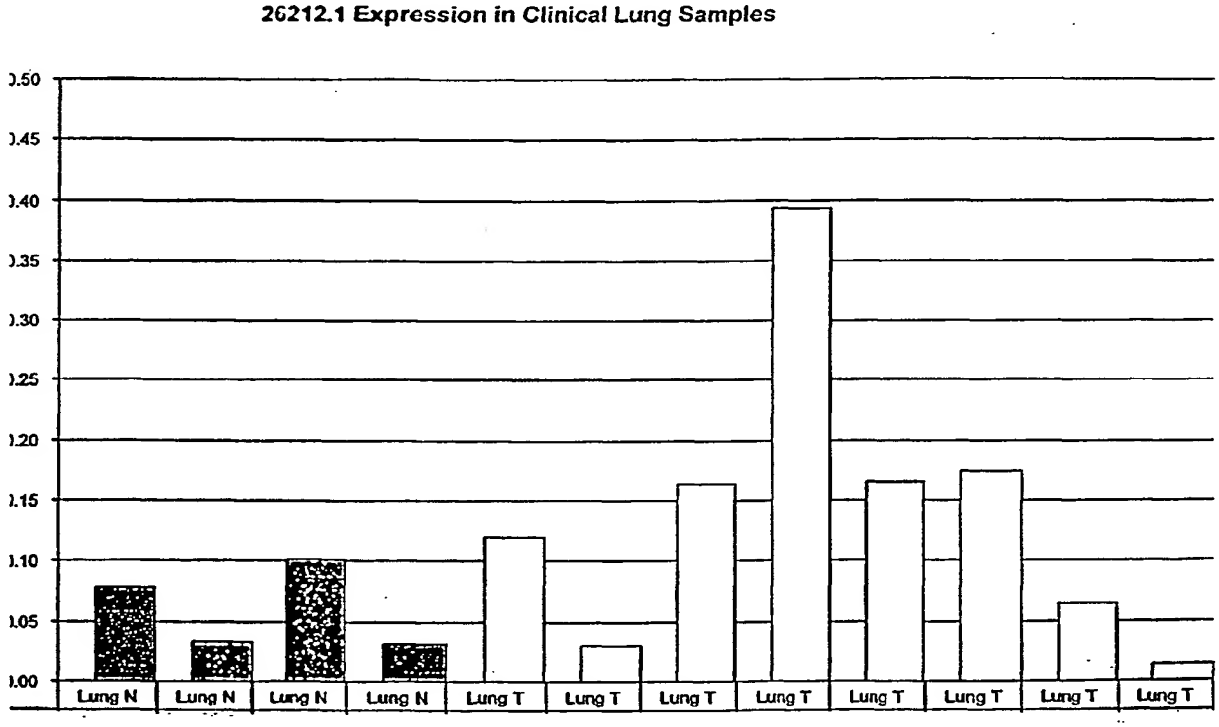

FIGURE 33

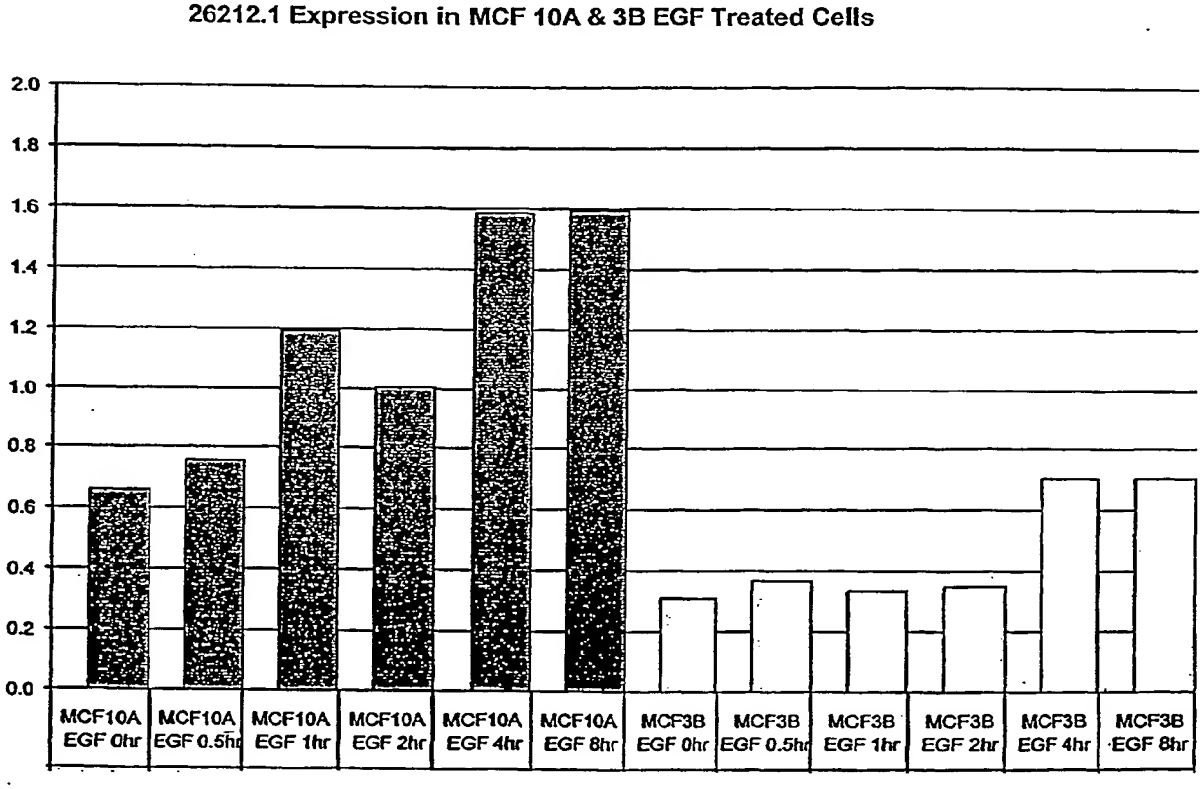
26212.1 Expression in MCF 10A & 3B EGF Treated Cells
2.0
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
MCF10A EGF 0hr
MCF10A EGF 0.5hr
MCF10A EGF 1hr
MCF10A EGF 2hr
MCF10A EGF 4hr
MCF10A EGF 8hr
MCF3B EGF 0hr
MCF3B EGF 0.5hr
MCF3B EGF 1hr
MCF3B EGF 2hr
MCF3B EGF 4hr
MCF3B EGF 8hr

FIGURE 34

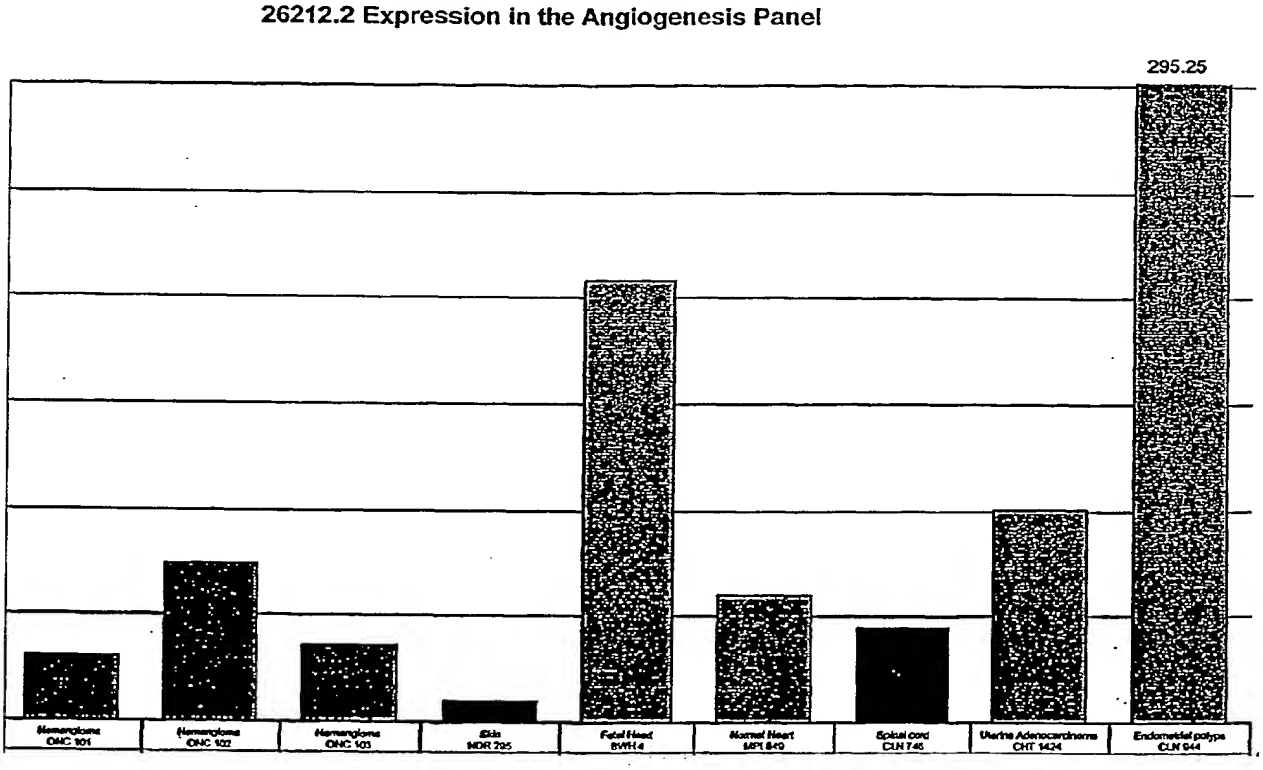

FIGURE 35

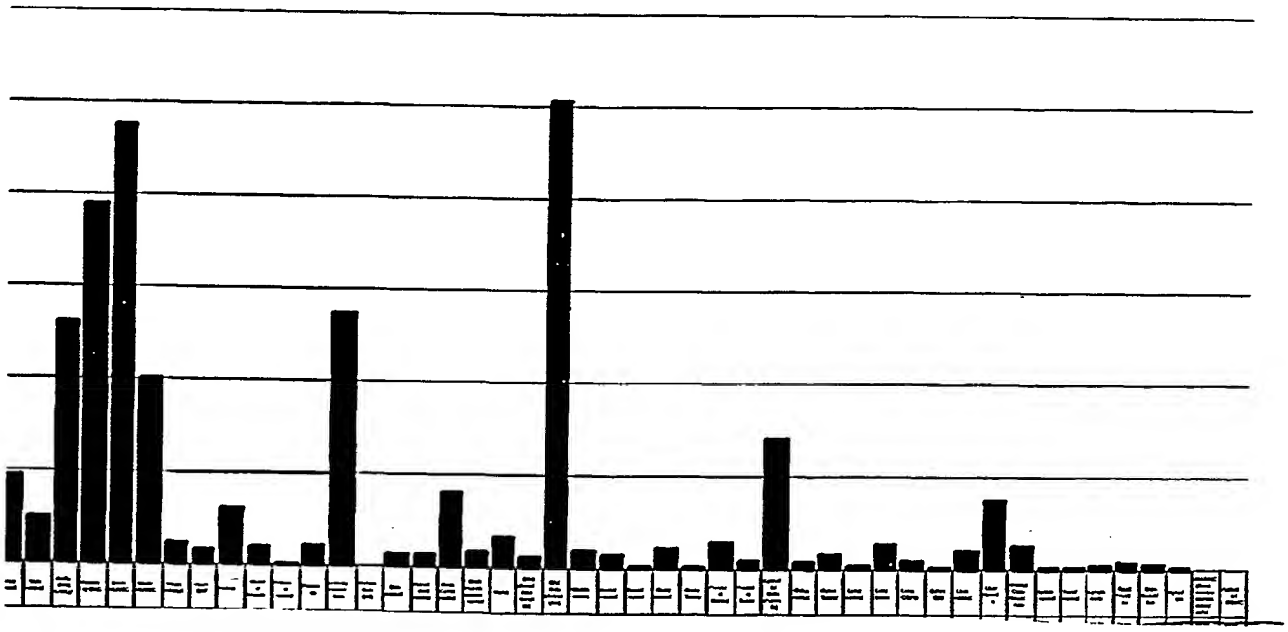

FIGURE 36

SEQUENCE LISTING

<110> Glucksman, Maria Alexandra
Williamson, Mark
Tsia, Fong-Ying
Rudolph-Owen, Laura A.

<120> 22438, 23553, 25278, and 26212 Novel Human Sulfatases (A CIP Application)

<130> 35800/208709

<151>

<160> 14

<170> FastSEQ for Windows Version 4.0

<210> 1
<211> 525
<212> PRT
<213> homo sapiens

<400> 1

```
Met Gly Trp Leu Phe Leu Lys Val Leu Leu Ala Gly Val Ser Phe Ser
 1               5                  10                  15
Gly Phe Leu Tyr Pro Leu Val Asp Phe Cys Ile Ser Gly Lys Thr Arg
            20                  25                  30
Gly Gln Lys Pro Asn Phe Val Ile Ile Leu Ala Asp Asp Met Gly Trp
        35                  40                  45
Gly Asp Leu Gly Ala Asn Trp Ala Glu Thr Lys Asp Thr Ala Asn Leu
    50                  55                  60
Asp Lys Met Ala Ser Glu Gly Met Arg Phe Val Asp Phe His Ala Ala
65                  70                  75                  80
Ala Ser Thr Cys Ser Pro Ser Arg Ala Ser Leu Leu Thr Gly Arg Leu
                85                  90                  95
Gly Leu Arg Asn Gly Val Thr Arg Asn Phe Ala Val Thr Ser Val Gly
            100                 105                 110
Gly Leu Pro Leu Asn Glu Thr Thr Leu Ala Glu Val Leu Gln Gln Ala
        115                 120                 125
Gly Tyr Val Thr Gly Ile Ile Gly Lys Trp His Leu Gly His His Gly
    130                 135                 140
Ser Tyr His Pro Asn Phe Arg Gly Phe Asp Tyr Tyr Phe Gly Ile Pro
145                 150                 155                 160
Tyr Ser His Asp Met Gly Cys Thr Asp Thr Pro Gly Tyr Asn His Pro
                165                 170                 175
Pro Cys Pro Ala Cys Pro Gln Gly Asp Gly Pro Ser Arg Asn Leu Gln
            180                 185                 190
Arg Asp Cys Tyr Thr Asp Val Ala Leu Pro Leu Tyr Glu Asn Leu Asn
        195                 200                 205
Ile Val Glu Gln Pro Val Asn Leu Ser Ser Leu Ala Gln Lys Tyr Ala
    210                 215                 220
Glu Lys Ala Thr Gln Phe Ile Gln Arg Ala Ser Thr Ser Gly Arg Pro
225                 230                 235                 240
Phe Leu Leu Tyr Val Ala Leu Ala His Met His Val Pro Leu Pro Val
                245                 250                 255
Thr Gln Leu Pro Ala Ala Pro Arg Gly Arg Ser Leu Tyr Gly Ala Gly
            260                 265                 270
Leu Trp Glu Met Asp Ser Leu Val Gly Gln Ile Lys Asp Lys Val Asp
        275                 280                 285
His Thr Val Lys Glu Asn Thr Phe Leu Trp Phe Thr Gly Asp Asn Gly
    290                 295                 300
Pro Trp Ala Gln Lys Cys Glu Leu Ala Gly Ser Val Gly Pro Phe Thr
305                 310                 315                 320
Gly Phe Trp Gln Thr Arg Gln Gly Gly Ser Pro Ala Lys Gln Thr Thr
```

```
                325                 330                 335
Trp Glu Gly Gly His Arg Val Pro Ala Leu Ala Tyr Trp Pro Gly Arg
            340                 345                 350
Val Pro Val Asn Val Thr Ser Thr Ala Leu Leu Ser Val Leu Asp Ile
        355                 360                 365
Phe Pro Thr Val Val Ala Leu Ala Gln Ala Ser Leu Pro Gln Gly Arg
    370                 375                 380
Arg Phe Asp Gly Val Asp Val Ser Glu Val Leu Phe Gly Arg Ser Gln
385                 390                 395                 400
Pro Gly His Arg Val Leu Phe His Pro Asn Ser Gly Ala Ala Gly Glu
                405                 410                 415
Phe Gly Ala Leu Gln Thr Val Arg Leu Glu Arg Tyr Lys Ala Phe Tyr
            420                 425                 430
Ile Thr Gly Gly Ala Arg Ala Cys Asp Gly Ser Thr Gly Pro Glu Leu
        435                 440                 445
Gln His Lys Phe Pro Leu Ile Phe Asn Leu Glu Asp Asp Thr Ala Glu
    450                 455                 460
Ala Val Pro Leu Glu Arg Gly Gly Ala Glu Tyr Gln Ala Val Leu Pro
465                 470                 475                 480
Glu Val Arg Lys Val Leu Ala Asp Val Leu Gln Asp Ile Ala Asn Asp
                485                 490                 495
Asn Ile Ser Ser Ala Asp Tyr Thr Gln Asp Pro Ser Val Thr Pro Cys
            500                 505                 510
Cys Asn Pro Tyr Gln Ile Ala Cys Arg Cys Gln Ala Ala
        515                 520                 525
```

<210> 2
<211> 2175
<212> DNA
<213> homo sapiens

<220>
<221> CDS
<222> (248)...(1825)

<400> 2

```
cacgcgtccg caaatttcct gattcttttg aattaggatt ccagatgggg gcctcatttc     60
tacagccccc aacattccta tagccgttat cactgccatc accactgcca ccagcatctt    120
cttgcagatt ccacccctgc tccccagaga cttcctgctt tgaaagtgag cagaaaggaa    180
gctctcagaa aaatctctag tggtggctgc cgtcgctcca gacaatcgga atcctgcctt    240
caccacc atg ggc tgg ctt ttt cta aag gtt ttg ttg gcg gga gtg agt       289
        Met Gly Trp Leu Phe Leu Lys Val Leu Leu Ala Gly Val Ser
         1               5                  10

ttc tca gga ttt ctt tat cct ctt gtg gat ttt tgc atc agt ggg aaa      337
Phe Ser Gly Phe Leu Tyr Pro Leu Val Asp Phe Cys Ile Ser Gly Lys
 15                  20                  25                  30

aca aga gga cag aag cca aac ttt gtg att att ttg gcc gat gac atg      385
Thr Arg Gly Gln Lys Pro Asn Phe Val Ile Ile Leu Ala Asp Asp Met
                 35                  40                  45

ggg tgg ggt gac ctg gga gca aac tgg gca gaa aca aag gac act gcc      433
Gly Trp Gly Asp Leu Gly Ala Asn Trp Ala Glu Thr Lys Asp Thr Ala
             50                  55                  60

aac ctt gat aag atg gct tcg gag gga atg agg ttt gtg gat ttc cat      481
Asn Leu Asp Lys Met Ala Ser Glu Gly Met Arg Phe Val Asp Phe His
         65                  70                  75

gca gct gcc tcc acc tgc tca ccc tcc cgg gct tcc ttg ctc acc ggc      529
Ala Ala Ala Ser Thr Cys Ser Pro Ser Arg Ala Ser Leu Leu Thr Gly
     80                  85                  90

cgg ctt ggc ctt cgc aat gga gtc aca cgc aac ttt gca gtc act tct      577
Arg Leu Gly Leu Arg Asn Gly Val Thr Arg Asn Phe Ala Val Thr Ser
 95                  100                 105                 110
```

```
gtg gga ggc ctt ccg ctc aac gag acc acc ttg gca gag gtg ctg cag      625
Val Gly Gly Leu Pro Leu Asn Glu Thr Thr Leu Ala Glu Val Leu Gln
                115                 120                 125

cag gcg ggt tac gtc act ggg ata ata ggc aaa tgg cat ctt gga cac      673
Gln Ala Gly Tyr Val Thr Gly Ile Ile Gly Lys Trp His Leu Gly His
            130                 135                 140

cac ggc tct tat cac ccc aac ttc cgt ggt ttt gat tac tac ttt gga      721
His Gly Ser Tyr His Pro Asn Phe Arg Gly Phe Asp Tyr Tyr Phe Gly
        145                 150                 155

atc cca tat agc cat gat atg ggc tgt act gat act cca ggc tac aac      769
Ile Pro Tyr Ser His Asp Met Gly Cys Thr Asp Thr Pro Gly Tyr Asn
    160                 165                 170

cac cct cct tgt cca gcg tgt cca cag ggt gat gga cca tca agg aac      817
His Pro Pro Cys Pro Ala Cys Pro Gln Gly Asp Gly Pro Ser Arg Asn
175                 180                 185                 190

ctt caa aga gac tgt tac act gac gtg gcc ctc cct ctt tat gaa aac      865
Leu Gln Arg Asp Cys Tyr Thr Asp Val Ala Leu Pro Leu Tyr Glu Asn
                195                 200                 205

ctc aac att gtg gag cag ccg gtg aac ttg agc agc ctt gcc cag aag      913
Leu Asn Ile Val Glu Gln Pro Val Asn Leu Ser Ser Leu Ala Gln Lys
            210                 215                 220

tat gct gag aaa gca acc cag ttc atc cag cgt gca agc acc agc ggg      961
Tyr Ala Glu Lys Ala Thr Gln Phe Ile Gln Arg Ala Ser Thr Ser Gly
        225                 230                 235

agg ccc ttc ctg ctc tat gtg gct ctg gcc cac atg cac gtg ccc tta     1009
Arg Pro Phe Leu Leu Tyr Val Ala Leu Ala His Met His Val Pro Leu
    240                 245                 250

ccc gtg act cag cta cca gca gcg cca cgg ggc aga agc ctg tat ggt     1057
Pro Val Thr Gln Leu Pro Ala Ala Pro Arg Gly Arg Ser Leu Tyr Gly
255                 260                 265                 270

gca ggg ctc tgg gag atg gac agt ctg gtg ggc cag atc aag gac aaa     1105
Ala Gly Leu Trp Glu Met Asp Ser Leu Val Gly Gln Ile Lys Asp Lys
                275                 280                 285

gtt gac cac aca gtg aag gaa aac aca ttc ctc tgg ttt aca gga gac     1153
Val Asp His Thr Val Lys Glu Asn Thr Phe Leu Trp Phe Thr Gly Asp
            290                 295                 300

aat ggc ccg tgg gct cag aag tgt gag cta gcg ggc agt gtg ggt ccc     1201
Asn Gly Pro Trp Ala Gln Lys Cys Glu Leu Ala Gly Ser Val Gly Pro
        305                 310                 315

ttc act gga ttt tgg caa act cgt caa ggg gga agt cca gcc aag cag     1249
Phe Thr Gly Phe Trp Gln Thr Arg Gln Gly Gly Ser Pro Ala Lys Gln
    320                 325                 330

acg acc tgg gaa gga ggg cac cgg gtc cca gca ctg gct tac tgg cct     1297
Thr Thr Trp Glu Gly Gly His Arg Val Pro Ala Leu Ala Tyr Trp Pro
335                 340                 345                 350

ggc aga gtt cca gtt aat gtc acc agc act gcc ttg tta agc gtg ctg     1345
Gly Arg Val Pro Val Asn Val Thr Ser Thr Ala Leu Leu Ser Val Leu
                355                 360                 365

gac att ttt cca act gtg gta gcc ctg gcc cag gcc agc tta cct caa     1393
Asp Ile Phe Pro Thr Val Val Ala Leu Ala Gln Ala Ser Leu Pro Gln
```

```
            370                 375                 380

gga cgg cgc ttt gat ggt gtg gac gtc tcc gag gtg ctc ttt ggc cgg     1441
Gly Arg Arg Phe Asp Gly Val Asp Val Ser Glu Val Leu Phe Gly Arg
        385                 390                 395

tca cag cct ggg cac agg gtg ctg ttc cac ccc aac agc ggg gca gct     1489
Ser Gln Pro Gly His Arg Val Leu Phe His Pro Asn Ser Gly Ala Ala
    400                 405                 410

gga gag ttt gga gcc ctg cag act gtc cgc ctg gag cgt tac aag gcc     1537
Gly Glu Phe Gly Ala Leu Gln Thr Val Arg Leu Glu Arg Tyr Lys Ala
415                 420                 425                 430

ttc tac att acc ggt gga gcc agg gcg tgt gat ggg agc acg ggg cct     1585
Phe Tyr Ile Thr Gly Gly Ala Arg Ala Cys Asp Gly Ser Thr Gly Pro
                435                 440                 445

gag ctg cag cat aag ttt cct ctg att ttc aac ctg gaa gac gat acc     1633
Glu Leu Gln His Lys Phe Pro Leu Ile Phe Asn Leu Glu Asp Asp Thr
            450                 455                 460

gca gaa gct gtg ccc cta gaa aga ggt ggt gcg gag tac cag gct gtg     1681
Ala Glu Ala Val Pro Leu Glu Arg Gly Gly Ala Glu Tyr Gln Ala Val
        465                 470                 475

ctg ccc gag gtc aga aag gtt ctt gca gac gtc ctc caa gac att gcc     1729
Leu Pro Glu Val Arg Lys Val Leu Ala Asp Val Leu Gln Asp Ile Ala
    480                 485                 490

aac gac aac atc tcc agc gca gat tac act cag gac cct tca gta act     1777
Asn Asp Asn Ile Ser Ser Ala Asp Tyr Thr Gln Asp Pro Ser Val Thr
495                 500                 505                 510

ccc tgc tgt aat ccc tac caa att gcc tgc cgc tgt caa gcc gca taa     1825
Pro Cys Cys Asn Pro Tyr Gln Ile Ala Cys Arg Cys Gln Ala Ala  * .
                515                 520                 525

cagaccaatt tttattccac gaggaggagt acctggaaat taggcaagtt tgcttccaaa   1885
tttcattttt accctcttta caaacacacg ctttagttta gtcttggagt ttagttttgg   1945
agttagcctt gcatatccct tctgtatcct gtccctcctc cacgccgacc cgagagcagc   2005
tgagctgcgc tggctctggg cagggagtgt gccttaatgg gaagcacacg ggctttggag   2065
tcaggcacag gtgccagctc cagcttttga acttgggcaa ttgtttaacc taacctgcaa   2125
gttgattttg agggttaaat aaaggcatac atgaaaaaaa aaaaaaaaaa              2175

<210> 3
<211> 871
<212> PRT
<213> homo sapiens

<400> 3
Met Lys Tyr Ser Cys Cys Ala Leu Val Leu Ala Val Leu Gly Thr Glu
 1               5                  10                  15
Leu Leu Gly Ser Leu Cys Ser Thr Val Arg Ser Pro Arg Phe Arg Gly
            20                  25                  30
Arg Ile Gln Gln Glu Arg Lys Asn Ile Arg Pro Asn Ile Ile Leu Val
        35                  40                  45
Leu Thr Asp Asp Gln Asp Val Glu Leu Gly Ser Leu Gln Val Met Asn
    50                  55                  60
Lys Thr Arg Lys Ile Met Glu His Gly Gly Ala Thr Phe Ile Asn Ala
65                  70                  75                  80
Phe Val Thr Thr Pro Met Cys Cys Pro Ser Arg Ser Ser Met Leu Thr
                85                  90                  95
Gly Lys Tyr Val His Asn His Asn Val Tyr Thr Asn Asn Glu Asn Cys
            100                 105                 110
Ser Ser Pro Ser Trp Gln Ala Met His Glu Pro Arg Thr Phe Ala Val
        115                 120                 125
```

```
Tyr Leu Asn Asn Thr Gly Tyr Arg Thr Ala Phe Phe Gly Lys Tyr Leu
    130                 135                 140
Asn Glu Tyr Asn Gly Ser Tyr Ile Pro Pro Gly Trp Arg Glu Trp Leu
145                 150                 155                 160
Gly Leu Ile Lys Asn Ser Arg Phe Tyr Asn Tyr Thr Val Cys Arg Asn
                165                 170                 175
Gly Ile Lys Glu Lys His Gly Phe Asp Tyr Ala Lys Asp Tyr Phe Thr
            180                 185                 190
Asp Leu Ile Thr Asn Glu Ser Ile Asn Tyr Phe Lys Met Ser Lys Arg
        195                 200                 205
Met Tyr Pro His Arg Pro Val Met Met Val Ile Ser His Ala Ala Pro
    210                 215                 220
His Gly Pro Glu Asp Ser Ala Pro Gln Phe Ser Lys Leu Tyr Pro Asn
225                 230                 235                 240
Ala Ser Gln His Ile Thr Pro Ser Tyr Asn Tyr Ala Pro Asn Met Asp
                245                 250                 255
Lys His Trp Ile Met Gln Tyr Thr Gly Pro Met Leu Pro Ile His Met
            260                 265                 270
Glu Phe Thr Asn Ile Leu Gln Arg Lys Arg Leu Gln Thr Leu Met Ser
        275                 280                 285
Val Asp Asp Ser Val Glu Arg Leu Tyr Asn Met Leu Val Glu Thr Gly
    290                 295                 300
Glu Leu Glu Asn Thr Tyr Ile Ile Tyr Thr Ala Asp His Gly Tyr His
305                 310                 315                 320
Ile Gly Gln Phe Gly Leu Val Lys Gly Lys Ser Met Pro Tyr Asp Phe
                325                 330                 335
Asp Ile Arg Val Pro Phe Phe Ile Arg Gly Pro Ser Val Glu Pro Gly
            340                 345                 350
Ser Ile Val Pro Gln Ile Val Leu Asn Ile Asp Leu Ala Pro Thr Ile
        355                 360                 365
Leu Asp Ile Ala Gly Leu Asp Thr Pro Pro Asp Val Asp Gly Lys Ser
    370                 375                 380
Val Leu Lys Leu Leu Asp Pro Glu Lys Pro Gly Asn Arg Phe Arg Thr
385                 390                 395                 400
Asn Lys Lys Ala Lys Ile Trp Arg Asp Thr Phe Leu Val Glu Arg Gly
                405                 410                 415
Lys Phe Leu Arg Lys Lys Glu Glu Ser Ser Lys Asn Ile Gln Gln Ser
            420                 425                 430
Asn His Leu Pro Lys Tyr Glu Arg Val Lys Glu Leu Cys Gln Gln Ala
        435                 440                 445
Arg Tyr Gln Thr Ala Cys Glu Gln Pro Gly Gln Lys Trp Gln Cys Ile
    450                 455                 460
Glu Asp Thr Ser Gly Lys Leu Arg Ile His Lys Cys Lys Gly Pro Ser
465                 470                 475                 480
Asp Leu Leu Thr Val Arg Gln Ser Thr Arg Asn Leu Tyr Ala Arg Gly
                485                 490                 495
Phe His Asp Lys Asp Lys Glu Cys Ser Cys Arg Glu Ser Gly Tyr Arg
            500                 505                 510
Ala Ser Arg Ser Gln Arg Lys Ser Gln Arg Gln Phe Leu Arg Asn Gln
        515                 520                 525
Gly Thr Pro Lys Tyr Lys Pro Arg Phe Val His Thr Arg Gln Thr Arg
    530                 535                 540
Ser Leu Ser Val Glu Phe Glu Gly Glu Ile Tyr Asp Ile Asn Leu Glu
545                 550                 555                 560
Glu Glu Glu Glu Leu Gln Val Leu Gln Pro Arg Asn Ile Ala Lys Arg
                565                 570                 575
His Asp Glu Gly His Lys Gly Pro Arg Asp Leu Gln Ala Ser Ser Gly
            580                 585                 590
Gly Asn Arg Gly Arg Met Leu Ala Asp Ser Ser Asn Ala Val Gly Pro
        595                 600                 605
Pro Thr Thr Val Arg Val Thr His Lys Cys Phe Ile Leu Pro Asn Asp
    610                 615                 620
Ser Ile His Cys Glu Arg Glu Leu Tyr Gln Ser Ala Arg Ala Trp Lys
625                 630                 635                 640
Asp His Lys Ala Tyr Ile Asp Lys Glu Ile Glu Ala Leu Gln Asp Lys
                645                 650                 655
Ile Lys Asn Leu Arg Glu Val Arg Gly His Leu Lys Arg Arg Lys Pro
```

```
            660                 665                 670
Glu Glu Cys Ser Cys Ser Lys Gln Ser Tyr Tyr Asn Lys Glu Lys Gly
        675                 680                 685
Val Lys Lys Gln Glu Lys Leu Lys Ser His Leu His Pro Phe Lys Glu
    690                 695                 700
Ala Ala Gln Glu Val Asp Ser Lys Leu Gln Leu Phe Lys Glu Asn Asn
705                 710                 715                 720
Arg Arg Arg Lys Lys Glu Arg Lys Glu Lys Arg Arg Gln Arg Lys Gly
                725                 730                 735
Glu Glu Cys Ser Leu Pro Gly Leu Thr Cys Phe Thr His Asp Asn Asn
            740                 745                 750
His Trp Gln Thr Ala Pro Phe Trp Asn Leu Gly Ser Phe Cys Ala Cys
        755                 760                 765
Thr Ser Ser Asn Asn Asn Thr Tyr Trp Cys Leu Arg Thr Val Asn Glu
    770                 775                 780
Thr His Asn Phe Leu Phe Cys Glu Phe Ala Thr Gly Phe Leu Glu Tyr
785                 790                 795                 800
Phe Asp Met Asn Thr Asp Pro Tyr Gln Leu Thr Asn Thr Val His Thr
                805                 810                 815
Val Glu Arg Gly Ile Leu Asn Gln Leu His Val Gln Leu Met Glu Leu
            820                 825                 830
Arg Ser Cys Gln Gly Tyr Lys Gln Cys Asn Pro Arg Pro Lys Asn Leu
        835                 840                 845
Asp Val Gly Asn Lys Asp Gly Gly Ser Tyr Asp Leu His Arg Gly Gln
    850                 855                 860
Leu Trp Asp Gly Trp Glu Gly
865                 870
```

<210> 4
<211> 4321
<212> DNA
<213> homo sapiens

<220>
<221> CDS
<222> (510)...(3125)

<400> 4

```
cccacgcgtc cggctaatga atcttggggc cggtgtcggg ccggggcggc ttgatcggca     60
actaggaaac cccaggcgca gaggccagga gcgagggcag cgaggatcag aggccaggcc    120
ttcccggctg ccggcgctcc tcggaggtca gggcagatga ggaacatgac tctccccctt    180
cggaggagga aggaagtccc gctgccacct tatctctgct cctctgcctc ctccctgttc    240
ccagagcttt ttctctagag aagattttga aggcggcttt tgtgctgacg gccacccacc    300
atcatctaaa gaagataaac ttggcaaatg acatgcaggt tcttcaaggc agaataattg    360
cagaaaatct tcaaaggacc ctatctgcag atgttctgaa tacctctgag aatagagatt    420
gattattcaa ccaggatacc taattcaaga actccagaaa tcaggagacg gagacatttt    480
gtcagttttg caacattgga ccaaataca atg aag tat tct tgc tgt gct ctg      533
                              Met Lys Tyr Ser Cys Cys Ala Leu
                               1               5

gtt ttg gct gtc ctg ggc aca gaa ttg ctg gga agc ctc tgt tcg act      581
Val Leu Ala Val Leu Gly Thr Glu Leu Leu Gly Ser Leu Cys Ser Thr
     10                  15                  20

gtc aga tcc ccg agg ttc aga gga cgg ata cag cag gaa cga aaa aac      629
Val Arg Ser Pro Arg Phe Arg Gly Arg Ile Gln Gln Glu Arg Lys Asn
 25                  30                  35                  40

atc cga ccc aac att att ctt gtg ctt acc gat gat caa gat gtg gag      677
Ile Arg Pro Asn Ile Ile Leu Val Leu Thr Asp Asp Gln Asp Val Glu
                 45                  50                  55

ctg ggg tcc ctg caa gtc atg aac aaa acg aga aag att atg gaa cat      725
Leu Gly Ser Leu Gln Val Met Asn Lys Thr Arg Lys Ile Met Glu His
             60                  65                  70

ggg ggg gcc acc ttc atc aat gcc ttt gtg act aca ccc atg tgc tgc      773
```

```
Gly Gly Ala Thr Phe Ile Asn Ala Phe Val Thr Thr Pro Met Cys Cys
          75                  80                  85

ccg tca cgg tcc tcc atg ctc acc ggg aag tat gtg cac aat cac aat      821
Pro Ser Arg Ser Ser Met Leu Thr Gly Lys Tyr Val His Asn His Asn
     90                  95                 100

gtc tac acc aac aac gag aac tgc tct tcc ccc tcg tgg cag gcc atg      869
Val Tyr Thr Asn Asn Glu Asn Cys Ser Ser Pro Ser Trp Gln Ala Met
105                 110                 115                 120

cat gag cct cgg act ttt gct gta tat ctt aac aac act ggc tac aga      917
His Glu Pro Arg Thr Phe Ala Val Tyr Leu Asn Asn Thr Gly Tyr Arg
                125                 130                 135

aca gcc ttt ttt gga aaa tac ctc aat gaa tat aat ggc agc tac atc      965
Thr Ala Phe Phe Gly Lys Tyr Leu Asn Glu Tyr Asn Gly Ser Tyr Ile
            140                 145                 150

ccc cct ggg tgg cga gaa tgg ctt gga tta atc aag aat tct cgc ttc     1013
Pro Pro Gly Trp Arg Glu Trp Leu Gly Leu Ile Lys Asn Ser Arg Phe
        155                 160                 165

tat aat tac act gtt tgt cgc aat ggc atc aaa gaa aag cat gga ttt     1061
Tyr Asn Tyr Thr Val Cys Arg Asn Gly Ile Lys Glu Lys His Gly Phe
    170                 175                 180

gat tat gca aag gac tac ttc aca gac tta atc act aac gag agc att     1109
Asp Tyr Ala Lys Asp Tyr Phe Thr Asp Leu Ile Thr Asn Glu Ser Ile
185                 190                 195                 200

aat tac ttc aaa atg tct aag aga atg tat ccc cat agg ccc gtt atg     1157
Asn Tyr Phe Lys Met Ser Lys Arg Met Tyr Pro His Arg Pro Val Met
                205                 210                 215

atg gtg atc agc cac gct gcg ccc cac ggc ccc gag gac tca gcc cca     1205
Met Val Ile Ser His Ala Ala Pro His Gly Pro Glu Asp Ser Ala Pro
            220                 225                 230

cag ttt tct aaa ctg tac ccc aat gct tcc caa cac ata act cct agt     1253
Gln Phe Ser Lys Leu Tyr Pro Asn Ala Ser Gln His Ile Thr Pro Ser
        235                 240                 245

tat aac tat gca cca aat atg gat aaa cac tgg att atg cag tac aca     1301
Tyr Asn Tyr Ala Pro Asn Met Asp Lys His Trp Ile Met Gln Tyr Thr
    250                 255                 260

gga cca atg ctg ccc atc cac atg gaa ttt aca aac att cta cag cgc     1349
Gly Pro Met Leu Pro Ile His Met Glu Phe Thr Asn Ile Leu Gln Arg
265                 270                 275                 280

aaa agg ctc cag act ttg atg tca gtg gat gat tct gtg gag agg ctg     1397
Lys Arg Leu Gln Thr Leu Met Ser Val Asp Asp Ser Val Glu Arg Leu
                285                 290                 295

tat aac atg ctc gtg gag acg ggg gag ctg gag aat act tac atc att     1445
Tyr Asn Met Leu Val Glu Thr Gly Glu Leu Glu Asn Thr Tyr Ile Ile
            300                 305                 310

tac acc gcc gac cat ggt tac cat att ggg cag ttt gga ctg gtc aag     1493
Tyr Thr Ala Asp His Gly Tyr His Ile Gly Gln Phe Gly Leu Val Lys
        315                 320                 325

ggg aaa tcc atg cca tat gac ttt gat att cgt gtg cct ttt ttt att     1541
Gly Lys Ser Met Pro Tyr Asp Phe Asp Ile Arg Val Pro Phe Phe Ile
    330                 335                 340
```

```
cgt ggt cca agt gta gaa cca gga tca ata gtc cca cag atc gtt ctc      1589
Arg Gly Pro Ser Val Glu Pro Gly Ser Ile Val Pro Gln Ile Val Leu
345                 350                 355                 360

aac att gac ttg gcc ccc acg atc ctg gat att gct ggg ctc gac aca      1637
Asn Ile Asp Leu Ala Pro Thr Ile Leu Asp Ile Ala Gly Leu Asp Thr
                365                 370                 375

cct cct gat gtg gac ggc aag tct gtc ctc aaa ctt ctg gac cca gaa      1685
Pro Pro Asp Val Asp Gly Lys Ser Val Leu Lys Leu Leu Asp Pro Glu
            380                 385                 390

aag cca ggt aac agg ttt cga aca aac aag aag gcc aaa att tgg cgt      1733
Lys Pro Gly Asn Arg Phe Arg Thr Asn Lys Lys Ala Lys Ile Trp Arg
        395                 400                 405

gat aca ttc cta gtg gaa aga ggc aaa ttt cta cgt aag aag gaa gaa      1781
Asp Thr Phe Leu Val Glu Arg Gly Lys Phe Leu Arg Lys Lys Glu Glu
    410                 415                 420

tcc agc aag aat atc caa cag tca aat cac ttg ccc aaa tat gaa cgg      1829
Ser Ser Lys Asn Ile Gln Gln Ser Asn His Leu Pro Lys Tyr Glu Arg
425                 430                 435                 440

gtc aaa gaa cta tgc cag cag gcc agg tac cag aca gcc tgt gaa caa      1877
Val Lys Glu Leu Cys Gln Gln Ala Arg Tyr Gln Thr Ala Cys Glu Gln
                445                 450                 455

ccg ggg cag aag tgg caa tgc att gag gat aca tct ggc aag ctt cga      1925
Pro Gly Gln Lys Trp Gln Cys Ile Glu Asp Thr Ser Gly Lys Leu Arg
            460                 465                 470

att cac aag tgt aaa gga ccc agt gac ctg ctc aca gtc cgg cag agc      1973
Ile His Lys Cys Lys Gly Pro Ser Asp Leu Leu Thr Val Arg Gln Ser
        475                 480                 485

acg cgg aac ctc tac gct cgc ggc ttc cat gac aaa gac aaa gag tgc      2021
Thr Arg Asn Leu Tyr Ala Arg Gly Phe His Asp Lys Asp Lys Glu Cys
    490                 495                 500

agt tgt agg gag tct ggt tac cgt gcc agc aga agc caa aga aag agt      2069
Ser Cys Arg Glu Ser Gly Tyr Arg Ala Ser Arg Ser Gln Arg Lys Ser
505                 510                 515                 520

caa cgg caa ttc ttg aga aac cag ggg act cca aag tac aag ccc aga      2117
Gln Arg Gln Phe Leu Arg Asn Gln Gly Thr Pro Lys Tyr Lys Pro Arg
                525                 530                 535

ttt gtc cat act cgg cag aca cgt tcc ttg tcc gtc gaa ttt gaa ggt      2165
Phe Val His Thr Arg Gln Thr Arg Ser Leu Ser Val Glu Phe Glu Gly
            540                 545                 550

gaa ata tat gac ata aat ctg gaa gaa gaa gaa gaa ttg caa gtg ttg      2213
Glu Ile Tyr Asp Ile Asn Leu Glu Glu Glu Glu Leu Gln Val Leu
        555                 560                 565

caa cca aga aac att gct aag cgt cat gat gaa ggc cac aag ggg cca      2261
Gln Pro Arg Asn Ile Ala Lys Arg His Asp Glu Gly His Lys Gly Pro
    570                 575                 580

aga gat ctc cag gct tcc agt ggt ggc aac agg ggc agg atg ctg gca      2309
Arg Asp Leu Gln Ala Ser Ser Gly Gly Asn Arg Gly Arg Met Leu Ala
585                 590                 595                 600

gat agc agc aac gcc gtg ggc cca cct acc act gtc cga gtg aca cac      2357
Asp Ser Ser Asn Ala Val Gly Pro Pro Thr Thr Val Arg Val Thr His
                605                 610                 615
```

```
aag tgt ttt att ctt ccc aat gac tct atc cat tgt gag aga gaa ctg      2405
Lys Cys Phe Ile Leu Pro Asn Asp Ser Ile His Cys Glu Arg Glu Leu
            620                 625                 630

tac caa tcg gcc aga gcg tgg aag gac cat aag gca tac att gac aaa      2453
Tyr Gln Ser Ala Arg Ala Trp Lys Asp His Lys Ala Tyr Ile Asp Lys
        635                 640                 645

gag att gaa gct ctg caa gat aaa att aag aat tta aga gaa gtg aga      2501
Glu Ile Glu Ala Leu Gln Asp Lys Ile Lys Asn Leu Arg Glu Val Arg
    650                 655                 660

gga cat ctg aag aga agg aag cct gag gaa tgt agc tgc agt aaa caa      2549
Gly His Leu Lys Arg Arg Lys Pro Glu Glu Cys Ser Cys Ser Lys Gln
665                 670                 675                 680

agc tat tac aat aaa gag aaa ggt gta aaa aag caa gag aaa tta aag      2597
Ser Tyr Tyr Asn Lys Glu Lys Gly Val Lys Lys Gln Glu Lys Leu Lys
                685                 690                 695

agc cat ctt cac cca ttc aag gag gct gct cag gaa gta gat agc aaa      2645
Ser His Leu His Pro Phe Lys Glu Ala Ala Gln Glu Val Asp Ser Lys
            700                 705                 710

ctg caa ctt ttc aag gag aac aac cgt agg agg aag aag gag agg aag      2693
Leu Gln Leu Phe Lys Glu Asn Asn Arg Arg Arg Lys Lys Glu Arg Lys
        715                 720                 725

gag aag aga cgg cag agg aag ggg gaa gag tgc agc ctg cct ggc ctc      2741
Glu Lys Arg Arg Gln Arg Lys Gly Glu Glu Cys Ser Leu Pro Gly Leu
    730                 735                 740

act tgc ttc acg cat gac aac aac cac tgg cag aca gcc ccg ttc tgg      2789
Thr Cys Phe Thr His Asp Asn Asn His Trp Gln Thr Ala Pro Phe Trp
745                 750                 755                 760

aac ctg gga tct ttc tgt gct tgc acg agt tct aac aat aac acc tac      2837
Asn Leu Gly Ser Phe Cys Ala Cys Thr Ser Ser Asn Asn Asn Thr Tyr
                765                 770                 775

tgg tgt ttg cgt aca gtt aat gag acg cat aat ttt ctt ttc tgt gag      2885
Trp Cys Leu Arg Thr Val Asn Glu Thr His Asn Phe Leu Phe Cys Glu
            780                 785                 790

ttt gct act ggc ttt ttg gag tat ttt gat atg aat aca gat cct tat      2933
Phe Ala Thr Gly Phe Leu Glu Tyr Phe Asp Met Asn Thr Asp Pro Tyr
        795                 800                 805

cag ctc aca aat aca gtg cac acg gta gaa cga ggc att ttg aat cag      2981
Gln Leu Thr Asn Thr Val His Thr Val Glu Arg Gly Ile Leu Asn Gln
    810                 815                 820

cta cac gta caa cta atg gag ctc aga agc tgt caa gga tat aag cag      3029
Leu His Val Gln Leu Met Glu Leu Arg Ser Cys Gln Gly Tyr Lys Gln
825                 830                 835                 840

tgc aac cca aga cct aag aat ctt gat gtt gga aat aaa gat gga gga      3077
Cys Asn Pro Arg Pro Lys Asn Leu Asp Val Gly Asn Lys Asp Gly Gly
                845                 850                 855

agc tat gac cta cac aga gga cag tta tgg gat gga tgg gaa ggt taa      3125
Ser Tyr Asp Leu His Arg Gly Gln Leu Trp Asp Gly Trp Glu Gly  *
            860                 865                 870

tcagccccgt ctcactgcag acatcaactg gcaaggccta gaggagctac acagtgtgaa   3185
tgaaaacatc tatgagtaca gacaaaacta cagacttagt ctggtggact ggactaatta   3245
```

cttgaaggat ttagatagag tatttgcact gctgaagagt cactatgagc aaaataaaac 3305
aaataagact caaactgctc aaagtgacgg gttcttggtt gtctctgctg agcacgctgt 3365
gtcaatggag atggcctctg ctgactcaga tgaagaccca aggcataagg ttgggaaaac 3425
acctcatttg accttgccag ctgaccttca aaccctgcat ttgaaccgac caacattaag 3485
tccagagagt aaacttgaat ggaataacga cattccagaa gttaatcatt tgaattctga 3545
acactggaga aaaaccgaaa aatggacggg gcatgaagag actaatcatc tggaaaccga 3605
tttcagtggc gatggcatga cagagctaga gctcgggccc agccccaggc tgcagcccat 3665
tcgcaggcac ccgaaagaac ttccccagta tggtggtcct ggaaaggaca tttttgaaga 3725
tcaactatat cttcctgtgc attccgatgg aatttcagtt catcagatgt tcaccatggc 3785
caccgcagaa caccgaagta attccagcat agcggggaag atgttgacca aggtggagaa 3845
gaatcacgaa aaggagaagt cacagcacct agaaggcagc gcctcctctt cactctcctc 3905
tgattagatg aaactgttac cttaccctaa acacagtatt tctttttaac tttttattt 3965
gtaaactaat aaaggkaatc acagccacca acattccaag ctaccctggg taccttttgtg 4025
cagtagaagc tagtgagcat gtgagcaagc ggtgtgcaca cggagactca tcgttataat 4085
ttactatctg ccaaggagta gaaagaaagg ctggggatat ttgggttggc tttggktttg 4145
atttttgct tggttggttg gtttgkacta aaacagtatt atcttttgaa tatcgtaggg 4205
acataarkww wwwmmwkktw wtcmawymra kakgsywrra wkgggstyty tskkrkstmw 4265
amwykwscmc cyskkrwwaw tywywmmywc mykytssstg rykrnktaat gaagtt 4321

<210> 5
<211> 569
<212> PRT
<213> homo sapiens

<400> 5
Met His Thr Leu Thr Gly Phe Ser Leu Val Ser Leu Leu Ser Phe Gly
1 5 10 15
Tyr Leu Ser Trp Asp Trp Ala Lys Pro Ser Phe Val Ala Asp Gly Pro
20 25 30
Gly Glu Ala Gly Glu Gln Pro Ser Ala Ala Pro Pro Gln Pro Pro His
35 40 45
Ile Ile Phe Ile Leu Thr Asp Asp Gln Gly Tyr His Asp Val Gly Tyr
50 55 60
His Gly Ser Asp Ile Glu Thr Pro Thr Leu Asp Arg Leu Ala Ala Lys
65 70 75 80
Gly Val Lys Leu Glu Asn Tyr Tyr Ile Gln Pro Ile Cys Thr Pro Ser
85 90 95
Arg Ser Gln Leu Leu Thr Gly Arg Tyr Gln Ile His Thr Gly Leu Gln
100 105 110
His Ser Ile Ile Arg Pro Gln Gln Pro Asn Cys Leu Pro Leu Asp Gln
115 120 125
Val Thr Leu Pro Gln Lys Leu Gln Glu Ala Gly Tyr Ser Thr His Met
130 135 140
Val Gly Lys Trp His Leu Gly Phe Tyr Arg Lys Glu Cys Leu Pro Thr
145 150 155 160
Arg Arg Gly Phe Asp Thr Phe Leu Gly Ser Leu Thr Gly Asn Val Asp
165 170 175
Tyr Tyr Thr Tyr Asp Asn Cys Asp Gly Pro Gly Val Cys Gly Phe Asp
180 185 190
Leu His Glu Gly Glu Asn Val Ala Trp Gly Leu Ser Gly Gln Tyr Ser
195 200 205
Thr Met Leu Tyr Ala Gln Arg Ala Ser His Ile Leu Ala Ser His Ser
210 215 220
Pro Gln Arg Pro Leu Phe Leu Tyr Val Ala Phe Gln Ala Val His Thr
225 230 235 240
Pro Leu Gln Ser Pro Arg Glu Tyr Leu Tyr Arg Tyr Arg Thr Met Gly
245 250 255
Asn Val Ala Arg Arg Lys Tyr Ala Ala Met Val Thr Cys Met Asp Glu
260 265 270
Ala Val Arg Asn Ile Thr Trp Ala Leu Lys Arg Tyr Gly Phe Tyr Asn
275 280 285
Asn Ser Val Ile Ile Phe Ser Ser Asp Asn Gly Gly Gln Thr Phe Ser
290 295 300
Gly Gly Ser Asn Trp Pro Leu Arg Gly Arg Lys Gly Thr Tyr Trp Glu
305 310 315 320
Gly Gly Val Arg Gly Leu Gly Phe Val His Ser Pro Leu Leu Lys Arg
325 330 335

```
Lys Gln Arg Thr Ser Arg Ala Leu Met His Ile Thr Asp Trp Tyr Pro
            340                 345                 350
Thr Leu Val Gly Leu Ala Gly Gly Thr Thr Ser Ala Ala Asp Gly Leu
        355                 360                 365
Asp Gly Tyr Asp Val Trp Pro Ala Ile Ser Glu Gly Arg Ala Ser Pro
    370                 375                 380
Arg Thr Glu Ile Leu His Asn Ile Asp Pro Leu Tyr Asn His Ala Gln
385                 390                 395                 400
His Gly Ser Leu Glu Gly Gly Phe Gly Ile Trp Asn Thr Ala Val Gln
                405                 410                 415
Ala Ala Ile Arg Val Gly Glu Trp Lys Leu Leu Thr Gly Asp Pro Gly
            420                 425                 430
Tyr Gly Asp Trp Ile Pro Pro Gln Thr Leu Ala Thr Phe Pro Gly Ser
        435                 440                 445
Trp Trp Asn Leu Glu Arg Met Ala Ser Val Arg Gln Ala Val Trp Leu
    450                 455                 460
Phe Asn Ile Ser Ala Asp Pro Tyr Glu Arg Glu Asp Leu Ala Gly Gln
465                 470                 475                 480
Arg Pro Asp Val Val Arg Thr Leu Leu Ala Arg Leu Ala Glu Tyr Asn
                485                 490                 495
Arg Thr Ala Ile Pro Val Arg Tyr Pro Ala Glu Asn Pro Arg Ala His
            500                 505                 510
Pro Asp Phe Asn Gly Gly Ala Trp Gly Pro Trp Ala Ser Asp Glu Glu
        515                 520                 525
Glu Glu Glu Glu Glu Gly Arg Ala Arg Ser Phe Ser Arg Gly Arg Arg
    530                 535                 540
Lys Lys Lys Cys Lys Ile Cys Lys Leu Arg Ser Phe Phe Arg Lys Leu
545                 550                 555                 560
Asn Thr Arg Leu Met Ser Gln Arg Ile
                565
```

<210> 6
<211> 2940
<212> DNA
<213> homo sapiens

<220>
<221> CDS
<222> (334)...(2043)

<400> 6

```
ccacgcgtcc gcccacgcgt ccggctgcca cgccgcgtct caggctggcc gggctgagcc     60
ggggaagagg gagcaaaggc ggcgcagggc ctgcgcttag gcagcgggag gcagctcggc    120
gcgggcctga cctccccaga gcgccccgct gcggccgagc agatccggcc cagccgtccg    180
gcagccagtc ccggaccaga cactggaccg tccccggggg gcgctgaact ccctcgcagc    240
atccgagccg gcgggccggt ggtgcgccct gggcgcgcga ggtggtgagg ccccaggagc    300
ccggcgcgcc gggacacgcg ggccggcttg gcg atg cac acc ctc act ggc ttc     354
                                     Met His Thr Leu Thr Gly Phe
                                      1               5

tct ctg gtc agc ctg ctc agc ttc ggc tac ctg tcc tgg gac tgg gcc      402
Ser Leu Val Ser Leu Leu Ser Phe Gly Tyr Leu Ser Trp Asp Trp Ala
         10                  15                  20

aag ccg agc ttc gtg gcc gac ggg ccc ggg gag gct ggc gag cag ccc      450
Lys Pro Ser Phe Val Ala Asp Gly Pro Gly Glu Ala Gly Glu Gln Pro
     25                  30                  35

tcg gcc gct ccg ccc cag cct ccc cac atc atc ttc atc ctc acg gac      498
Ser Ala Ala Pro Pro Gln Pro Pro His Ile Ile Phe Ile Leu Thr Asp
 40                  45                  50                  55

gac caa ggc tac cac gac gtg ggc tac cat ggt tca gat atc gag acc      546
Asp Gln Gly Tyr His Asp Val Gly Tyr His Gly Ser Asp Ile Glu Thr
                 60                  65                  70

cct acg ctg gac agg ctg gcg gcc aag ggg gtc aag ttg gag aat tat      594
```

```
Pro Thr Leu Asp Arg Leu Ala Ala Lys Gly Val Lys Leu Glu Asn Tyr
                75                  80                  85

tac atc cag ccc atc tgc acg cct tcg cgg agc cag ctc ctc act ggc      642
Tyr Ile Gln Pro Ile Cys Thr Pro Ser Arg Ser Gln Leu Leu Thr Gly
            90                  95                 100

agg tac cag atc cac aca gga ctc cag cat tcc atc atc cgc cca cag      690
Arg Tyr Gln Ile His Thr Gly Leu Gln His Ser Ile Ile Arg Pro Gln
        105                 110                 115

cag ccc aac tgc ctg ccc ctg gac cag gtg aca ctg cca cag aag ctg      738
Gln Pro Asn Cys Leu Pro Leu Asp Gln Val Thr Leu Pro Gln Lys Leu
    120                 125                 130                 135

cag gag gca ggt tat tcc acc cat atg gtg ggc aag tgg cac ctg ggc      786
Gln Glu Ala Gly Tyr Ser Thr His Met Val Gly Lys Trp His Leu Gly
                140                 145                 150

ttc tac cgg aag gag tgt ctg ccc acc cgt cgg ggc ttc gac acc ttc      834
Phe Tyr Arg Lys Glu Cys Leu Pro Thr Arg Arg Gly Phe Asp Thr Phe
            155                 160                 165

ctg ggc tcg ctc acg ggc aat gtg gac tat tac acc tat gac aac tgt      882
Leu Gly Ser Leu Thr Gly Asn Val Asp Tyr Tyr Thr Tyr Asp Asn Cys
        170                 175                 180

gat ggc cca ggc gtg tgc ggc ttc gac ctg cac gag ggt gag aat gtg      930
Asp Gly Pro Gly Val Cys Gly Phe Asp Leu His Glu Gly Glu Asn Val
    185                 190                 195

gcc tgg ggg ctc agc ggc cag tac tcc act atg ctt tac gcc cag cgc      978
Ala Trp Gly Leu Ser Gly Gln Tyr Ser Thr Met Leu Tyr Ala Gln Arg
200                 205                 210                 215

gcc agc cat atc ctg gcc agc cac agc cct cag cgt ccc ctc ttc ctc     1026
Ala Ser His Ile Leu Ala Ser His Ser Pro Gln Arg Pro Leu Phe Leu
                220                 225                 230

tat gtg gcc ttc cag gca gta cac aca ccc ctg cag tcc cct cgt gag     1074
Tyr Val Ala Phe Gln Ala Val His Thr Pro Leu Gln Ser Pro Arg Glu
            235                 240                 245

tac ctg tac cgc tac cgc acc atg ggc aat gtg gcc cgg cgg aag tac     1122
Tyr Leu Tyr Arg Tyr Arg Thr Met Gly Asn Val Ala Arg Arg Lys Tyr
        250                 255                 260

gcg gcc atg gtg acc tgc atg gat gag gct gtg cgc aac atc acc tgg     1170
Ala Ala Met Val Thr Cys Met Asp Glu Ala Val Arg Asn Ile Thr Trp
    265                 270                 275

gcc ctc aag cgc tac ggt ttc tac aac aac agt gtc atc atc ttc tcc     1218
Ala Leu Lys Arg Tyr Gly Phe Tyr Asn Asn Ser Val Ile Ile Phe Ser
280                 285                 290                 295

agt gac aat ggt ggc cag act ttc tcg ggg ggc agc aac tgg ccg ctc     1266
Ser Asp Asn Gly Gly Gln Thr Phe Ser Gly Gly Ser Asn Trp Pro Leu
                300                 305                 310

cga gga cgc aag ggc act tat tgg gaa ggt ggc gtg cgg ggc cta ggc     1314
Arg Gly Arg Lys Gly Thr Tyr Trp Glu Gly Gly Val Arg Gly Leu Gly
            315                 320                 325

ttt gtc cac agt ccc ctg ctc aag cga aag caa cgg aca agc cgg gca     1362
Phe Val His Ser Pro Leu Leu Lys Arg Lys Gln Arg Thr Ser Arg Ala
        330                 335                 340
```

```
ctg atg cac atc act gac tgg tac ccg acc ctg gtg ggt ctg gca ggt     1410
Leu Met His Ile Thr Asp Trp Tyr Pro Thr Leu Val Gly Leu Ala Gly
    345                 350                 355

ggt acc acc tca gca gcc gat ggg cta gat ggc tac gac gtg tgg ccg     1458
Gly Thr Thr Ser Ala Ala Asp Gly Leu Asp Gly Tyr Asp Val Trp Pro
360                 365                 370                 375

gcc atc agc gag ggc cgg gcc tca cca cgc acg gag atc ctg cac aac     1506
Ala Ile Ser Glu Gly Arg Ala Ser Pro Arg Thr Glu Ile Leu His Asn
                380                 385                 390

att gac cca ctc tac aac cat gcc cag cat ggc tcc ctg gag ggc ggc     1554
Ile Asp Pro Leu Tyr Asn His Ala Gln His Gly Ser Leu Glu Gly Gly
            395                 400                 405

ttt ggc atc tgg aac acc gcc gtg cag gct gcc atc cgc gtg ggt gag     1602
Phe Gly Ile Trp Asn Thr Ala Val Gln Ala Ala Ile Arg Val Gly Glu
        410                 415                 420

tgg aag ctg ctg aca gga gac ccc ggc tat ggc gat tgg atc cca ccg     1650
Trp Lys Leu Leu Thr Gly Asp Pro Gly Tyr Gly Asp Trp Ile Pro Pro
    425                 430                 435

cag aca ctg gcc acc ttc ccg ggt agc tgg tgg aac ctg gaa cga atg     1698
Gln Thr Leu Ala Thr Phe Pro Gly Ser Trp Trp Asn Leu Glu Arg Met
440                 445                 450                 455

gcc agt gtc cgc cag gcc gtg tgg ctc ttc aac atc agt gct gac cct     1746
Ala Ser Val Arg Gln Ala Val Trp Leu Phe Asn Ile Ser Ala Asp Pro
                460                 465                 470

tat gaa cgg gag gac ctg gct ggc cag cgg cct gat gtg gtc cgc acc     1794
Tyr Glu Arg Glu Asp Leu Ala Gly Gln Arg Pro Asp Val Val Arg Thr
            475                 480                 485

ctg ctg gct cgc ctg gcc gaa tat aac cgc aca gcc atc ccg gta cgc     1842
Leu Leu Ala Arg Leu Ala Glu Tyr Asn Arg Thr Ala Ile Pro Val Arg
        490                 495                 500

tac cca gct gag aac ccc cgg gct cat cct gac ttt aat ggg ggt gct     1890
Tyr Pro Ala Glu Asn Pro Arg Ala His Pro Asp Phe Asn Gly Gly Ala
    505                 510                 515

tgg ggg ccc tgg gcc agt gat gag gaa gag gag gaa gag gaa ggg agg     1938
Trp Gly Pro Trp Ala Ser Asp Glu Glu Glu Glu Glu Glu Glu Gly Arg
520                 525                 530                 535

gct cga agc ttc tcc cgg ggt cgt cgc aag aaa aaa tgc aag att tgc     1986
Ala Arg Ser Phe Ser Arg Gly Arg Arg Lys Lys Lys Cys Lys Ile Cys
                540                 545                 550

aag ctt cga tcc ttt ttc cgt aaa ctc aac acc agg cta atg tcc caa     2034
Lys Leu Arg Ser Phe Phe Arg Lys Leu Asn Thr Arg Leu Met Ser Gln
            555                 560                 565

cgg atc tga tggtggggag ggagaaaact gtcctttaga ggatcttccc             2083
Arg Ile  *


cactccggct tggccctgct gtttctcagg gagaagcctg tcacatctcc atctacaggg   2143
agttggaggg tgtagagtcc cttggttgaa cagggtaggg agcctggata ggagtgggtg   2203
ggaataaacc agactgggat gcctgtgtct cagtcctgcc tcctcacgga cttgctctgt   2263
gacctcaggt gacccacatg agcttttagc ctcagtttcc tcatctgtaa aatgagctct   2323
aatgactttg tgactctttg gtgtggccct ggagcctggg gccacggtgg agttcctggc   2383
cggccttgcc acttgacaac tcctttaagg cttccccctt aacacgggat ccctgtggtg   2443
gtgtttggga gttgcctgga ggcaactcca agcctggccc ccagctgaag catggcaatc   2503
```

```
tggctgctct ctacagggac ccccaagcgc tgtgggtgga gggcaggggt cgggggggtt   2563
gaccttcttg ggtcttcaca tggcctaggc cagtcctccg gtcagactgg tgtcaggcac   2623
cgtggtgcaa aattcctctt ctggcccctc cagtacccag agaaactggc tgggccatta   2683
actgctgcag caccaagggt ggtagaaaga gctgtgaaga gcccccaaac cagtaccagg   2743
acacctgggt tctcctgtga cctggggcac agttcttgcc ctctaggcct tgatttcccc   2803
acctgcaagt ggggatgcca gccctggctc tgcctccttc atgaggctct ggaagactgg   2863
ccaaggttgt ggaggagctt gtgaacttga ttaaagtgtc gtaacatgga aaaaaaaaaa   2923
aaaaaaaaaa agggcgg                                                  2940
```

<210> 7
<211> 599
<212> PRT
<213> homo sapiens

<400> 7

```
Met Ala Pro Arg Gly Cys Ala Gly His Pro Pro Pro Pro Ser Pro Gln
 1               5                  10                  15
Ala Cys Val Cys Pro Gly Lys Met Leu Ala Met Gly Ala Leu Ala Gly
            20                  25                  30
Phe Trp Ile Leu Cys Leu Leu Thr Tyr Gly Tyr Leu Ser Trp Gly Gln
        35                  40                  45
Ala Leu Glu Glu Glu Glu Glu Gly Ala Leu Leu Ala Gln Ala Gly Glu
    50                  55                  60
Lys Leu Glu Pro Ser Thr Thr Ser Thr Ser Gln Pro His Leu Ile Phe
65                  70                  75                  80
Ile Leu Ala Asp Asp Gln Gly Phe Arg Asp Val Gly Tyr His Gly Ser
                85                  90                  95
Glu Ile Lys Thr Pro Thr Leu Asp Lys Leu Ala Ala Glu Gly Val Lys
            100                 105                 110
Leu Glu Asn Tyr Tyr Val Gln Pro Ile Cys Thr Pro Ser Arg Ser Gln
        115                 120                 125
Phe Ile Thr Gly Lys Tyr Gln Ile His Thr Gly Leu Gln His Ser Ile
    130                 135                 140
Ile Arg Pro Thr Gln Pro Asn Cys Leu Pro Leu Asp Asn Ala Thr Leu
145                 150                 155                 160
Pro Gln Lys Leu Lys Glu Val Gly Tyr Ser Thr His Met Val Gly Lys
                165                 170                 175
Trp His Leu Gly Phe Tyr Arg Lys Glu Cys Met Pro Thr Arg Arg Gly
            180                 185                 190
Phe Asp Thr Phe Phe Gly Ser Leu Leu Gly Ser Gly Asp Tyr Tyr Thr
        195                 200                 205
His Tyr Lys Cys Asp Ser Pro Gly Met Cys Gly Tyr Asp Leu Tyr Glu
    210                 215                 220
Asn Asp Asn Ala Ala Trp Asp Tyr Asp Asn Gly Ile Tyr Ser Thr Gln
225                 230                 235                 240
Met Tyr Thr Gln Arg Val Gln Gln Ile Leu Ala Ser His Asn Pro Thr
                245                 250                 255
Lys Pro Ile Phe Leu Tyr Ile Ala Tyr Gln Ala Val His Ser Pro Leu
            260                 265                 270
Gln Ala Pro Gly Arg Tyr Phe Glu His Tyr Arg Ser Ile Ile Asn Ile
        275                 280                 285
Asn Arg Arg Arg Tyr Ala Ala Met Leu Ser Cys Leu Asp Glu Ala Ile
    290                 295                 300
Asn Asn Val Thr Leu Ala Leu Lys Thr Tyr Gly Phe Tyr Asn Asn Ser
305                 310                 315                 320
Ile Ile Ile Tyr Ser Ser Asp Asn Gly Gly Gln Pro Thr Ala Gly Gly
                325                 330                 335
Ser Asn Trp Pro Leu Arg Gly Ser Lys Gly Thr Tyr Trp Glu Gly Gly
            340                 345                 350
Ile Arg Ala Val Gly Phe Val His Ser Pro Leu Leu Lys Asn Lys Gly
        355                 360                 365
Thr Val Cys Lys Glu Leu Val His Ile Thr Asp Trp Tyr Pro Thr Leu
    370                 375                 380
Ile Ser Leu Ala Glu Gly Gln Ile Asp Glu Asp Ile Gln Leu Asp Gly
385                 390                 395                 400
Tyr Asp Ile Trp Glu Thr Ile Ser Glu Gly Leu Arg Ser Pro Arg Val
                405                 410                 415
```

```
Asp Ile Leu His Asn Ile Asp Pro Ile Tyr Thr Lys Ala Lys Asn Gly
            420                 425                 430
Ser Trp Ala Ala Gly Tyr Gly Ile Trp Asn Thr Ala Ile Gln Ser Ala
        435                 440                 445
Ile Arg Val Gln His Trp Lys Leu Leu Thr Gly Asn Pro Gly Tyr Ser
    450                 455                 460
Asp Trp Val Pro Pro Gln Ser Phe Ser Asn Leu Gly Pro Asn Arg Trp
465                 470                 475                 480
His Asn Glu Arg Ile Thr Leu Ser Thr Gly Lys Ser Val Trp Leu Phe
                485                 490                 495
Asn Ile Thr Ala Asp Pro Tyr Glu Arg Val Asp Leu Ser Asn Arg Tyr
            500                 505                 510
Pro Gly Ile Val Lys Lys Leu Leu Arg Arg Leu Ser Gln Phe Asn Lys
        515                 520                 525
Thr Ala Val Pro Val Arg Tyr Pro Pro Lys Asp Pro Arg Ser Asn Pro
    530                 535                 540
Arg Leu Asn Gly Gly Val Trp Gly Pro Trp Tyr Lys Glu Glu Thr Lys
545                 550                 555                 560
Lys Lys Lys Pro Ser Lys Asn Gln Ala Glu Lys Lys Gln Lys Lys Ser
                565                 570                 575
Lys Lys Lys Lys Lys Lys Gln Gln Lys Ala Val Ser Gly Ser Thr Cys
            580                 585                 590
His Ser Gly Val Thr Cys Gly
        595
```

<210> 8
<211> 2253
<212> DNA
<213> homo sapiens

<220>
<221> CDS
<222> (324)...(2123)

<400> 8

```
cacgcgtccg cccacgcgtc cgtggagata ttaacttttt tctttttttt tttccttggt      60
ggaagctgct ctagggaggg gggaggagga ggagaaagtg aaatgtgctg gagaagagcg     120
agccctcctt gttcttccgg agtcccatcc attaagccat cacttctgga agattaaagt     180
tgtcggacat ggtgacagct gagaggagag gaggatttct tgccaggtgg agagtcttca     240
ccgtctgttg ggtgcatgtg tgcgcccgca gcggcgcggg gcgcgtggtt ctccgcgtgg     300
agtctcacct gggacctgag tga atg gct ccc agg ggc tgt gcg ggg cat ccg      353
                         Met Ala Pro Arg Gly Cys Ala Gly His Pro
                          1               5                  10

cct ccg cct tct cca cag gcc tgt gtc tgt cct gga aag atg cta gca      401
Pro Pro Pro Ser Pro Gln Ala Cys Val Cys Pro Gly Lys Met Leu Ala
                15                  20                  25

atg ggg gcg ctg gca gga ttc tgg atc ctc tgc ctc ctc act tat ggt      449
Met Gly Ala Leu Ala Gly Phe Trp Ile Leu Cys Leu Leu Thr Tyr Gly
            30                  35                  40

tac ctg tcc tgg ggc cag gcc tta gaa gag gag gaa gaa ggg gcc tta      497
Tyr Leu Ser Trp Gly Gln Ala Leu Glu Glu Glu Glu Glu Gly Ala Leu
        45                  50                  55

cta gct caa gct gga gag aaa cta gag ccc agc aca act tcc acc tcc      545
Leu Ala Gln Ala Gly Glu Lys Leu Glu Pro Ser Thr Thr Ser Thr Ser
    60                  65                  70

cag ccc cat ctc att ttc atc cta gcg gat gat cag gga ttt aga gat      593
Gln Pro His Leu Ile Phe Ile Leu Ala Asp Asp Gln Gly Phe Arg Asp
75                  80                  85                  90

gtg ggt tac cac gga tct gag att aaa aca cct act ctt gac aag ctc      641
Val Gly Tyr His Gly Ser Glu Ile Lys Thr Pro Thr Leu Asp Lys Leu
                95                  100                 105
```

```
gct gcc gaa gga gtt aaa ctg gag aac tac tat gtc cag cct att tgc      689
Ala Ala Glu Gly Val Lys Leu Glu Asn Tyr Tyr Val Gln Pro Ile Cys
            110                 115                 120

aca cca tcc agg agt cag ttt att act gga aag tat cag ata cac acc      737
Thr Pro Ser Arg Ser Gln Phe Ile Thr Gly Lys Tyr Gln Ile His Thr
        125                 130                 135

gga ctt caa cat tct atc ata aga cct acc caa ccc aac tgt tta cct      785
Gly Leu Gln His Ser Ile Ile Arg Pro Thr Gln Pro Asn Cys Leu Pro
    140                 145                 150

ctg gac aat gcc acc cta cct cag aaa ctg aag gag gtt gga tat tca      833
Leu Asp Asn Ala Thr Leu Pro Gln Lys Leu Lys Glu Val Gly Tyr Ser
155                 160                 165                 170

acg cat atg gtc gga aaa tgg cac ttg ggt ttt tac aga aaa gaa tgc      881
Thr His Met Val Gly Lys Trp His Leu Gly Phe Tyr Arg Lys Glu Cys
                175                 180                 185

atg ccc acc aga aga gga ttt gat acc ttt ttt ggt tcc ctt ttg gga      929
Met Pro Thr Arg Arg Gly Phe Asp Thr Phe Phe Gly Ser Leu Leu Gly
            190                 195                 200

agt ggg gat tac tat aca cac tac aaa tgt gac agt cct ggg atg tgt      977
Ser Gly Asp Tyr Tyr Thr His Tyr Lys Cys Asp Ser Pro Gly Met Cys
        205                 210                 215

ggc tat gac ttg tat gaa aac gac aat gct gcc tgg gac tat gac aat     1025
Gly Tyr Asp Leu Tyr Glu Asn Asp Asn Ala Ala Trp Asp Tyr Asp Asn
    220                 225                 230

ggc ata tac tcc aca cag atg tac act cag aga gta cag caa atc tta     1073
Gly Ile Tyr Ser Thr Gln Met Tyr Thr Gln Arg Val Gln Gln Ile Leu
235                 240                 245                 250

gct tcc cat aac ccc aca aag cct ata ttt tta tat att gcc tat caa     1121
Ala Ser His Asn Pro Thr Lys Pro Ile Phe Leu Tyr Ile Ala Tyr Gln
                255                 260                 265

gct gtt cat tca cca ctg caa gct cct ggc agg tat ttc gaa cac tac     1169
Ala Val His Ser Pro Leu Gln Ala Pro Gly Arg Tyr Phe Glu His Tyr
            270                 275                 280

cga tcc att atc aac ata aac agg agg aga tat gct gcc atg ctt tcc     1217
Arg Ser Ile Ile Asn Ile Asn Arg Arg Arg Tyr Ala Ala Met Leu Ser
        285                 290                 295

tgc tta gat gaa gca atc aac aac gtg aca ttg gct cta aag act tat     1265
Cys Leu Asp Glu Ala Ile Asn Asn Val Thr Leu Ala Leu Lys Thr Tyr
    300                 305                 310

ggt ttc tat aac aac agc att atc att tac tct tca gat aat ggt ggc     1313
Gly Phe Tyr Asn Asn Ser Ile Ile Ile Tyr Ser Ser Asp Asn Gly Gly
315                 320                 325                 330

cag cct acg gca gga ggg agt aac tgg cct ctc aga ggt agc aaa gga     1361
Gln Pro Thr Ala Gly Gly Ser Asn Trp Pro Leu Arg Gly Ser Lys Gly
                335                 340                 345

aca tat tgg gaa gga ggg atc cgg gct gta ggc ttt gtg cat agc cca     1409
Thr Tyr Trp Glu Gly Gly Ile Arg Ala Val Gly Phe Val His Ser Pro
            350                 355                 360

ctt ctg aaa aac aag gga aca gtg tgt aag gaa ctt gtg cac atc act     1457
Leu Leu Lys Asn Lys Gly Thr Val Cys Lys Glu Leu Val His Ile Thr
```

```
         365                 370                 375

gac tgg tac ccc act ctc att tca ctg gct gaa gga cag att gat gag     1505
Asp Trp Tyr Pro Thr Leu Ile Ser Leu Ala Glu Gly Gln Ile Asp Glu
     380                 385                 390

gac att caa cta gat ggc tat gat atc tgg gag acc ata agt gag ggt     1553
Asp Ile Gln Leu Asp Gly Tyr Asp Ile Trp Glu Thr Ile Ser Glu Gly
395                 400                 405                 410

ctt cgc tca ccc cga gta gat att ttg cat aac att gac ccc ata tac     1601
Leu Arg Ser Pro Arg Val Asp Ile Leu His Asn Ile Asp Pro Ile Tyr
                415                 420                 425

acc aag gca aaa aat ggc tcc tgg gca gca ggc tat ggg atc tgg aac     1649
Thr Lys Ala Lys Asn Gly Ser Trp Ala Ala Gly Tyr Gly Ile Trp Asn
            430                 435                 440

act gca atc cag tca gcc atc aga gtg cag cac tgg aaa ttg ctt aca     1697
Thr Ala Ile Gln Ser Ala Ile Arg Val Gln His Trp Lys Leu Leu Thr
        445                 450                 455

gga aat cct ggc tac agc gac tgg gtc ccc cct cag tct ttc agc aac     1745
Gly Asn Pro Gly Tyr Ser Asp Trp Val Pro Pro Gln Ser Phe Ser Asn
    460                 465                 470

ctg gga ccg aac cgg tgg cac aat gaa cgg atc acc ttg tca act ggc     1793
Leu Gly Pro Asn Arg Trp His Asn Glu Arg Ile Thr Leu Ser Thr Gly
475                 480                 485                 490

aaa agt gta tgg ctt ttc aac atc aca gcc gac cca tat gag agg gtg     1841
Lys Ser Val Trp Leu Phe Asn Ile Thr Ala Asp Pro Tyr Glu Arg Val
                495                 500                 505

gac cta tct aac agg tat cca gga atc gtg aag aag ctc cta cgg agg     1889
Asp Leu Ser Asn Arg Tyr Pro Gly Ile Val Lys Lys Leu Leu Arg Arg
            510                 515                 520

ctc tca cag ttc aac aaa act gca gtg ccg gtc agg tat ccc ccc aaa     1937
Leu Ser Gln Phe Asn Lys Thr Ala Val Pro Val Arg Tyr Pro Pro Lys
        525                 530                 535

gac ccc aga agt aac cct agg ctc aat gga ggg gtc tgg gga cca tgg     1985
Asp Pro Arg Ser Asn Pro Arg Leu Asn Gly Gly Val Trp Gly Pro Trp
    540                 545                 550

tat aaa gag gaa acc aag aaa aag aag cca agc aaa aat cag gct gag     2033
Tyr Lys Glu Glu Thr Lys Lys Lys Lys Pro Ser Lys Asn Gln Ala Glu
555                 560                 565                 570

aaa aag caa aag aaa agc aaa aaa aag aag aag aaa cag cag aaa gca     2081
Lys Lys Gln Lys Lys Ser Lys Lys Lys Lys Lys Lys Gln Gln Lys Ala
                575                 580                 585

gtc tca ggt tca act tgc cat tca ggt gtt act tgt gga taa             2123
Val Ser Gly Ser Thr Cys His Ser Gly Val Thr Cys Gly  *
            590                 595

gcacaaatat ttcctgtttg gttaaacttt aatcagttct tatctttcat ctgtttccta   2183
ggtaaaccag caaatttggc tcgataatat cgctggccta agcgtcaggc ttgttttcat   2243
gctgtgccac                                                         2253

<210> 9
<211> 552
<212> PRT
<213> Artificial Sequence
```

<220>
<223> Pfam consensus sequence for human sulfatase

<400> 9

```
Pro Asn Ile Leu Leu Ile Leu Ala Asp Asp Leu Gly Ile Gly Asp Leu
 1               5                  10                  15
Gly Cys Tyr Gly Asn Pro Thr Ile Arg Thr Pro Asn Ile Asp Arg Leu
            20                  25                  30
Ala Glu Glu Gly Leu Arg Phe Thr Asn Ala Tyr Val Thr Thr Pro Leu
        35                  40                  45
Cys Thr Pro Ser Arg Ala Ala Leu Leu Thr Gly Arg Tyr Pro His Arg
    50                  55                  60
Thr Gly Met Tyr Thr Asn Asn Arg Ala Gly Val Leu Pro Phe Thr Gly
65                  70                  75                  80
Trp Ser Leu Glu Gly Gly Leu Pro Leu Asp Glu Thr Thr Leu Pro Glu
                85                  90                  95
Leu Leu Lys Glu Ala Gly Tyr Ala Thr Gly Met Val Gly Lys Trp His
            100                 105                 110
Gly Tyr Asn Glu Glu Ser Ser Ala Ser Asp Phe Ala His Leu Pro Leu
        115                 120                 125
Gly Arg Gly Phe Asp Tyr Phe Tyr Gly Asn Leu Gly Gly Glu Asp Gln
    130                 135                 140
Trp Tyr Pro Leu Val Asp Ala Leu Leu Pro Phe Thr Asn Asp Thr Tyr
145                 150                 155                 160
Thr Cys Glu Gly Gly Tyr Gly Phe Ser Lys Asp Val Ala Leu Lys Pro
                165                 170                 175
Leu Gly Ala Leu Gly Val Asn Glu Val Glu Ala Pro Asp Lys Ala Leu
            180                 185                 190
Ala Asp Tyr Lys Thr Ala Gly Ala Leu Asn Val Pro His His Val Phe
        195                 200                 205
Glu Trp Ala Asp Arg Tyr Ala Gly Ala Val Asp Val Gly Arg Pro Phe
    210                 215                 220
Leu Ala Val Leu Ile Phe Pro Arg Pro Ala Ala Cys Phe Leu Tyr Pro
225                 230                 235                 240
Asn Ala Thr Val Val Ser Gln Pro Met Pro His Ser Pro Leu Thr Ala
                245                 250                 255
Pro Arg Pro Trp Gln Leu Leu Ala Asp Glu Ala Leu Pro Phe Leu Glu
            260                 265                 270
Arg Asn Gly Gln Arg Asp Lys Pro Phe Phe Leu Tyr Leu Ser Tyr Lys
        275                 280                 285
His Val His Ile Pro Arg Asp Ala Pro Met Leu Phe Ser Ser Lys Asp
    290                 295                 300
Phe Ala Gly Ser Ser Arg Arg Gly Leu Tyr Gly Leu Ile Leu Asp Ser
305                 310                 315                 320
Val Glu Glu Met Asp Asp Gly Val Gly Arg Val Leu Asn Ala Leu Asp
                325                 330                 335
Glu Leu Asn Gly Leu Leu Asp Asn Thr Leu Ile Ile Phe Thr Ser Leu
            340                 345                 350
Leu Asp His Gly Gly His Leu Gly Ala His Gly His Leu Gly Ile Arg
        355                 360                 365
Ala Gly Gly Ser Asn Gly Pro Phe Arg Gly Gly Lys Gly Thr Asn Leu
    370                 375                 380
Tyr Glu Gly Gly Thr Arg Val Pro Leu Ile Val Arg Trp Pro Glu Gly
385                 390                 395                 400
Ile Ile Ala Pro Gly Gln Val Ser Asp Glu Leu Val Ser Leu Met Asp
                405                 410                 415
Leu Phe Pro Thr Ile Leu Asp Leu Ala Gly Ala Pro Leu Pro Gly Val
            420                 425                 430
Ala Ala Gly Val Lys Asp Arg Ile Leu Asp Gly Val Ser Leu Leu Pro
        435                 440                 445
Leu Leu Leu Gly Ala Ala Gly Ser Ser Arg His Glu Thr Leu Phe Tyr
    450                 455                 460
Glu Ser Tyr Cys Asn Glu Gly Arg Gly Phe Leu Pro Ala Val Arg Trp
465                 470                 475                 480
Gly Lys Lys Lys Ala His Phe Arg Thr Pro Asn Ile Ala Gly Trp Gln
                485                 490                 495
Arg Val Asp Phe Asp Asp Val Trp Lys Leu Phe Asn Thr Val Glu Asp
```

```
            500                 505                 510
Phe Asn Arg Ser Gly Asp Asp Ala Cys Arg His Gly Asp Val Cys Lys
        515                 520                 525
Cys Leu Gly Lys Pro Arg Arg Ser Val Thr His His Asp Pro Pro Leu
    530                 535                 540
Leu Tyr Asp Leu Ser Arg Asp Pro
545                 550
```

<210> 10
<211> 520
<212> PRT
<213> Artificial Sequence

<220>
<223> Pfam consensus sequence for human sulfatase

<400> 10

```
Pro Asn Val Leu Leu Ile Leu Ala Asp Asp Leu Gly Ile Gly Asp Leu
 1               5                  10                  15
Gly Cys Tyr Gly His Pro Thr Ile Arg Thr Pro Asn Leu Asp Arg Leu
            20                  25                  30
Ala Glu Glu Gly Leu Arg Phe Thr Asn His Tyr Thr Ala Thr Pro Leu
        35                  40                  45
Cys Ser Pro Ser Arg Ala Ala Leu Leu Thr Gly Arg Tyr Pro His Arg
    50                  55                  60
His Gly Met Val Ser Asn Gly Arg Leu Gly Val Leu Gly Phe Thr Ala
65                  70                  75                  80
Lys Ser Gly Gly Leu Pro Leu Asp Glu Thr Thr Leu Pro Glu Leu Leu
                85                  90                  95
Lys Glu Ala Gly Tyr Ala Thr Gly Leu Val Gly Lys Trp His Leu Gly
            100                 105                 110
Leu Asn Glu Asn Ser Asp Ala Ala Gly Asp Gly Glu His Leu Pro Leu
        115                 120                 125
Gly Trp Arg Gly Phe Asp Tyr Phe Asp Gly Phe Leu Tyr Gly Ser Pro
    130                 135                 140
Phe Thr Tyr Asp Glu Glu Asn Cys Asp Asn Gly Glu Gly Thr Glu Pro
145                 150                 155                 160
Pro Glu Ala Tyr Pro Glu Gln Gly Trp Leu Pro Gln Ile Leu Gly Tyr
                165                 170                 175
Tyr Leu Thr Asp Leu Leu Ala Asp Lys Ala Leu Gly Leu Leu Asp Val
            180                 185                 190
Ala Ser Ala Ala Gly Arg Leu Leu Ala Lys Ala Leu Ala Ala Ser Arg
        195                 200                 205
Pro Phe Phe Leu Tyr Ile Ser Pro Pro Ala Pro His Phe Ser Ile Leu
    210                 215                 220
Phe Arg Asn Phe Lys Glu Val Ala Gln Pro Tyr Arg Ala Pro Gln Leu
225                 230                 235                 240
Thr Gln Leu Phe Val Asp Glu Ala Ala Asp Phe Ile Glu Arg Asn Lys
                245                 250                 255
Glu Lys Pro Phe Phe Leu Tyr Leu Ala Phe Leu Arg Leu His Val His
            260                 265                 270
Thr Pro Leu Phe Ser Pro Ala Glu Asp Leu Glu Ser Lys Asp Phe Leu
        275                 280                 285
Gly Arg Ser Gln Arg Gly Arg Tyr Gly Asp Leu Val Glu Glu Met Asp
    290                 295                 300
Asp Leu Val Gly Arg Val Leu Asp Ala Leu Glu Asp Leu Gly Leu Leu
305                 310                 315                 320
Asp Asn Thr Leu Val Ile Phe Thr Ser Asp Asn Gly Ala His Leu Glu
                325                 330                 335
Gly Thr Pro Glu Trp Tyr Gly Gly Gly Asn Gly Pro Leu Lys Gly Gly
            340                 345                 350
Lys Gly Tyr Gly Ser Leu Tyr Glu Gly Gly Ile Arg Val Pro Leu Leu
        355                 360                 365
Val Arg Trp Pro Gly Gly Ile Ala Pro Ala Gly Arg Val Lys Glu Lys
    370                 375                 380
Ser Glu Leu Val Ser His Val Asp Leu Ala Pro Thr Ile Leu Asp Leu
385                 390                 395                 400
```

```
Ala Gly Ala Pro Leu Pro Lys Val Ala Asn Gly Ala Lys Asp Arg Pro
                405                 410                 415
Leu Asp Gly Val Ser Leu Leu Pro Leu Leu Leu Gly Gly Ala Ala Pro
            420                 425                 430
Ser Arg Arg Ala His Glu Thr Leu Phe His Tyr Asn Gly Lys Gly Arg
        435                 440                 445
Lys Leu Arg Ala Val Arg Trp Pro Arg Lys Ser Gly Lys Thr Pro Lys
    450                 455                 460
Leu Lys Ala His Phe Phe Thr Pro Ala Phe Asp Asp Asp Thr Asn Asn
465                 470                 475                 480
Gly Trp Glu Cys Val Gly Thr Val Ser Gln Ala Asp Asp Ile Glu Asp
                485                 490                 495
Cys Arg Cys Glu Gly Val Glu Thr Val Thr His His Asp Pro Pro Glu
            500                 505                 510
Leu Tyr Asp Leu Ser Arg Asp Pro
        515                 520
```

<210> 11
<211> 1578
<212> DNA
<213> homo sapiens

<400> 11

```
atgggctggc tttttctaaa ggttttgttg gcgggagtga gtttctcagg atttctttat      60
cctcttgtgg atttttgcat cagtgggaaa acaagaggac agaagccaaa ctttgtgatt     120
attttggccg atgacatggg gtggggtgac ctgggagcaa actgggcaga aacaaaggac     180
actgccaacc ttgataagat ggcttcggag ggaatgaggt ttgtggattt ccatgcagct     240
gcctccacct gctcaccctc ccgggcttcc ttgctcaccg gccggcttgg ccttcgcaat     300
ggagtcacac gcaactttgc agtcacttct gtgggaggcc ttccgctcaa cgagaccacc     360
ttggcagagg tgctgcagca ggcgggttac gtcactggga taataggcaa atggcatctt     420
ggacaccacg gctcttatca ccccaacttc cgtggttttg attactactt tggaatccca     480
tatagccatg atatgggctg tactgatact ccaggctaca accaccctcc ttgtccagcg     540
tgtccacagg gtgatggacc atcaaggaac cttcaaagag actgttacac tgacgtggcc     600
ctccctcttt atgaaaacct caacattgtg gagcagccgg tgaacttgag cagccttgcc     660
cagaagtatg ctgagaaagc aacccagttc atccagcgtg caagcaccag cgggaggccc     720
ttcctgctct atgtggctct ggcccacatg cacgtgccct tacccgtgac tcagctacca     780
gcagcgccac ggggcagaag cctgtatggt gcagggctct gggagatgga cagtctggtg     840
ggccagatca aggacaaagt tgaccacaca gtgaaggaaa acacattcct ctggtttaca     900
ggagacaatg gcccgtgggc tcagaagtgt gagctagcgg gcagtgtggg tcccttcact     960
ggattttggc aaactcgtca agggggaagt ccagccaagc agacgacctg ggaaggaggg    1020
caccgggtcc cagcactggc ttactggcct ggcagagttc cagttaatgt caccagcact    1080
gccttgttaa gcgtgctgga catttttcca actgtggtag ccctggccca ggccagctta    1140
cctcaaggac ggcgctttga tggtgtggac gtctccgagg tgctctttgg ccggtcacag    1200
cctgggcaca gggtgctgtt ccaccccaac agcggggcag ctggagagtt tggagccctg    1260
cagactgtcc gcctggagcg ttacaaggcc ttctacatta ccggtggagc cagggcgtgt    1320
gatgggagca cggggcctga gctgcagcat aagtttcctc tgattttcaa cctggaagac    1380
gataccgcag aagctgtgcc cctagaaaga ggtggtgcgg agtaccaggc tgtgctgccc    1440
gaggtcagaa aggttcttgc agacgtcctc caagacattg ccaacgacaa catctccagc    1500
gcagattaca ctcaggaccc ttcagtaact ccctgctgta atccctacca aattgcctgc    1560
cgctgtcaag ccgcataa                                                  1578
```

<210> 12
<211> 2616
<212> DNA
<213> homo sapiens

<400> 12

```
atgaagtatt cttgctgtgc tctggttttg gctgtcctgg gcacagaatt gctgggaagc      60
ctctgttcga ctgtcagatc cccgaggttc agaggacgga tacagcagga acgaaaaaac     120
atccgaccca acattattct tgtgcttacc gatgatcaag atgtggagct ggggtccctg     180
caagtcatga acaaaacgag aaagattatg gaacatgggg gggccacctt catcaatgcc     240
tttgtgacta cacccatgtg ctgcccgtca cggtcctcca tgctcaccgg gaagtatgtg     300
cacaatcaca atgtctacac caacaacgag aactgctctt ccccctcgtg gcaggccatg     360
catgagcctc ggacttttgc tgtatatctt aacaacactg gctacagaac agccttttct     420
ggaaaatacc tcaatgaata taatggcagc tacatccccc ctgggtggcg agaatggctt     480
ggattaatca agaattctcg cttctataat tacactgttt gtcgcaatgg catcaaagaa     540
aagcatggat ttgattatgc aaaggactac ttcacagact taatcactaa cgagagcatt     600
```

aattacttca aaatgtctaa gagaatgtat ccccataggc ccgttatgat ggtgatcagc 660
cacgctgcgc cccacggccc cgaggactca gccccacagt tttctaaact gtaccccaat 720
gcttcccaac acataactcc tagttataac tatgcaccaa atatggataa acactggatt 780
atgcagtaca caggaccaat gctgcccatc cacatggaat ttacaaacat tctacagcgc 840
aaaaggctcc agactttgat gtcagtggat gattctgtgg agaggctgta taacatgctc 900
gtggagacgg gggagctgga gaatacttac atcatttaca ccgccgacca tggttaccat 960
attgggcagt ttggactggt caaggggaaa tccatgccat atgactttga tattcgtgtg 1020
ccttttttta ttcgtggtcc aagtgtagaa ccaggatcaa tagtcccaca gatcgttctc 1080
aacattgact tggcccccac gatcctggat attgctgggc tcgacacacc tcctgatgtg 1140
gacggcaagt ctgtcctcaa acttctggac ccagaaaagc caggtaacag gtttcgaaca 1200
aacaagaagg ccaaaatttg gcgtgataca ttcctagtgg aaagaggcaa atttctacgt 1260
aagaaggaag aatccagcaa gaatatccaa cagtcaaatc acttgcccaa atatgaacgg 1320
gtcaaagaac tatgccagca ggccaggtac cagacagcct gtgaacaacc ggggcagaag 1380
tggcaatgca ttgaggatac atctggcaag cttcgaattc acaagtgtaa aggacccagt 1440
gacctgctca cagtccggca gagcacgcgg aacctctacg ctcgcggctt ccatgacaaa 1500
gacaaagagt gcagttgtag ggagtctggt taccgtgcca gcagaagcca aagaaagagt 1560
caacggcaat tcttgagaaa ccaggggact ccaaagtaca agcccagatt tgtccatact 1620
cggcagacac gttccttgtc cgtcgaattt gaaggtgaaa tatatgacat aaatctggaa 1680
gaagaagaag aattgcaagt gttgcaacca agaaacattg ctaagcgtca tgatgaaggc 1740
cacaaggggc caagagatct ccaggcttcc agtggtggca acaggggcag gatgctggca 1800
gatagcagca acgccgtggg cccacctacc actgtccgag tgacacacaa gtgttttatt 1860
cttcccaatg actctatcca ttgtgagaga gaactgtacc aatcggccag agcgtggaag 1920
gaccataagg catacattga caaagagatt gaagctctgc aagataaaat taagaattta 1980
agagaagtga gaggacatct gaagagaagg aagcctgagg aatgtagctg cagtaaacaa 2040
agctattaca ataaagagaa aggtgtaaaa aagcaagaga aattaaagag ccatcttcac 2100
ccattcaagg aggctgctca ggaagtagat agcaaactgc aacttttcaa ggagaacaac 2160
cgtaggagga agaaggagag gaaggagaag agacggcaga ggaaggggga agagtgcagc 2220
ctgcctggcc tcacttgctt cacgcatgac aacaaccact ggcagacagc cccgttctgg 2280
aacctgggat ctttctgtgc ttgcacgagt tctaacaata acacctactg gtgtttgcgt 2340
acagttaatg agacgcataa ttttcttttc tgtgagtttg ctactggctt tttggagtat 2400
tttgatatga atacagatcc ttatcagctc acaaatacag tgcacacggt agaacgaggc 2460
attttgaatc agctacacgt acaactaatg gagctcagaa gctgtcaagg atataagcag 2520
tgcaacccaa gacctaagaa tcttgatgtt ggaaataaag atggaggaag ctatgaccta 2580
cacagaggac agttatggga tggatgggaa ggttaa 2616

<210> 13
<211> 1710
<212> DNA
<213> homo sapiens

<400> 13
atgcacaccc tcactggctt ctctctggtc agcctgctca gcttcggcta cctgtcctgg 60
gactgggcca agccgagctt cgtggccgac gggcccgggg aggctggcga gcagccctcg 120
gccgctccgc cccagcctcc ccacatcatc ttcatcctca cggacgacca aggctaccac 180
gacgtgggct accatggttc agatatcgag acccctacgc tggacaggct ggcggccaag 240
ggggtcaagt tggagaatta ttacatccag cccatctgca cgccttcgcg gagccagctc 300
ctcactggca ggtaccagat ccacacagga ctccagcatt ccatcatccg cccacagcag 360
cccaactgcc tgcccctgga ccaggtgaca ctgccacaga agctgcagga ggcaggttat 420
tccacccata tggtgggcaa gtggcacctg ggcttctacc ggaaggagtg tctgcccacc 480
cgtcgggggct tcgacacctt cctgggctcg ctcacgggca atgtggacta ttacacctat 540
gacaactgtg atggcccagg cgtgtgcggc ttcgacctgc acgagggtga gaatgtggcc 600
tgggggctca gcggccagta ctccactatg ctttacgccc agcgcgccag ccatatcctg 660
gccagccaca gccctcagcg tcccctcttc ctctatgtgg ccttccaggc agtacacaca 720
cccctgcagt cccctcgtga gtacctgtac cgctaccgca ccatgggcaa tgtggcccgg 780
cggaagtacg cggccatggt gacctgcatg gatgaggctg tgcgcaacat cacctgggcc 840
ctcaagcgct acggtttcta caacaacagt gtcatcatct tctccagtga caatggtggc 900
cagactttct cgggggggcag caactggccg ctccgaggac gcaagggcac ttattgggaa 960
ggtggcgtgc ggggcctagg ctttgtccac agtcccctgc tcaagcgaaa gcaacggaca 1020
agccgggcac tgatgcacat cactgactgg tacccgaccc tggtgggtct ggcaggtggt 1080
accacctcag cagccgatgg gctagatggc tacgacgtgt ggccggccat cagcgagggc 1140
cgggcctcac cacgcacgga gatcctgcac aacattgacc cactctacaa ccatgcccag 1200
catggctccc tggagggcgg ctttggcatc tggaacaccg ccgtgcaggc tgccatccgc 1260
gtgggtgagt ggaagctgct gacaggagac cccggctatg gcgattggat cccaccgcag 1320
acactggcca ccttcccggg tagctggtgg aacctggaac gaatggccag tgtccgccag 1380
gccgtgtggc tcttcaacat cagtgctgac ccttatgaac gggaggacct ggctggccag 1440
cggcctgatg tggtccgcac cctgctggct cgcctggccg aatataaccg cacagccatc 1500
ccggtacgct acccagctga gaacccccgg gctcatcctg actttaatgg gggtgcttgg 1560

gggccctggg ccagtgatga ggaagaggag gaagaggaag ggagggctcg aagcttctcc 1620
cggggtcgtc gcaagaaaaa atgcaagatt tgcaagcttc gatccttttt ccgtaaactc 1680
aacaccaggc taatgtccca acggatctga 1710

<210> 14
<211> 1800
<212> DNA
<213> homo sapiens

<400> 14
atggctccca ggggctgtgc ggggcatccg cctccgcctt ctccacaggc ctgtgtctgt 60
cctggaaaga tgctagcaat gggggcgctg gcaggattct ggatcctctg cctcctcact 120
tatggttacc tgtcctgggg ccaggcctta gaagaggagg aagaaggggc cttactagct 180
caagctggag agaaactaga gcccagcaca acttccacct cccagcccca tctcattttc 240
atcctagcgg atgatcaggg atttagagat gtgggttacc acggatctga gattaaaaca 300
cctactcttg acaagctcgc tgccgaagga gttaaactgg agaactacta tgtccagcct 360
atttgcacac catccaggag tcagtttatt actggaaagt atcagataca caccggactt 420
caacattcta tcataagacc tacccaaccc aactgtttac ctctggacaa tgccacccta 480
cctcagaaac tgaaggaggt tggatattca acgcatatgg tcggaaaatg gcacttgggt 540
ttttacagaa aagaatgcat gcccaccaga agaggatttg atacctttt tggttccctt 600
ttgggaagtg gggattacta tacacactac aaatgtgaca gtcctgggat gtgtggctat 660
gacttgtatg aaaacgacaa tgctgcctgg gactatgaca atggcatata ctccacacag 720
atgtacactc agagagtaca gcaaatctta gcttcccata accccacaaa gcctatattt 780
ttatatattg cctatcaagc tgttcattca ccactgcaag ctcctggcag gtatttcgaa 840
cactaccgat ccattatcaa cataaacagg aggagatatg ctgccatgct ttcctgctta 900
gatgaagcaa tcaacaacgt gacattggct ctaaagactt atggtttcta taacaacagc 960
attatcattt actcttcaga taatggtggc cagcctacgg caggagggag taactggcct 1020
ctcagaggta gcaaaggaac atattgggaa ggagggatcc gggctgtagg ctttgtgcat 1080
agcccacttc tgaaaaacaa gggaacagtg tgtaaggaac ttgtgcacat cactgactgg 1140
taccccactc tcatttcact ggctgaagga cagattgatg aggacattca actagatggc 1200
tatgatatct gggagaccat aagtgagggt cttcgctcac cccgagtaga tattttgcat 1260
aacattgacc ccatatacac caaggcaaaa aatggctcct gggcagcagg ctatgggatc 1320
tggaacactg caatccagtc agccatcaga gtgcagcact ggaaattgct tacaggaaat 1380
cctggctaca gcgactgggt ccccctcag tctttcagca acctgggacc gaaccggtgg 1440
cacaatgaac ggatcacctt gtcaactggc aaaagtgtat ggcttttcaa catcacagcc 1500
gacccatatg agagggtgga cctatctaac aggtatccag gaatcgtgaa gaagctccta 1560
cggaggctct cacagttcaa caaaactgca gtgccggtca ggtatccccc caaagacccc 1620
agaagtaacc ctaggctcaa tggaggggtc tggggaccat ggtataaaga ggaaaccaag 1680
aaaaagaagc caagcaaaaa tcaggctgag aaaaagcaaa agaaaagcaa aaaaaagaag 1740
aagaaacagc agaaagcagt ctcaggttca acttgccatt caggtgttac ttgtggataa 1800